

# OpsEval: A Comprehensive Task-Oriented AIOps Benchmark for Large Language Models

Yuhe Liu  
Tsinghua University  
Beijing, China

Bohan Chen  
Tsinghua University  
Beijing, China

Yongqian Sun  
Nankai University  
Tianjin, China

Haiming Zhang  
Chinese Academy of Sciences  
Beijing, China

Xidao Wen  
BizSeer  
Beijing, China

Changhua Pei  
Chinese Academy of Sciences  
Beijing, China

Mingze Sun  
Tsinghua University  
Beijing, China

Shenglin Zhang  
Nankai University  
Beijing, China

Jianhui Li  
Chinese Academy of Sciences  
Beijing, China

Xiaohui Nie  
BizSeer  
Beijing, China

Longlong Xu  
Tsinghua University  
Beijing, China

Zhirui Zhang  
Beijing University of Posts and  
Telecommunications  
Beijing, China

Kun Wang  
Tsinghua University  
Beijing, China

Gaogang Xie  
Chinese Academy of Sciences  
Beijing, China

Dan Pei\*  
Tsinghua University  
Beijing, China

## ABSTRACT

Large language models (LLMs) have exhibited remarkable capabilities in NLP-related tasks such as translation, summarizing, and generation. The application of LLMs in specific areas, notably AIOps (Artificial Intelligence for IT Operations), holds great potential due to their advanced abilities in information summarizing, report analyzing, and ability of API calling. Nevertheless, the performance of current LLMs in AIOps tasks is yet to be determined. Furthermore, a comprehensive benchmark is required to steer the optimization of LLMs tailored for AIOps. Compared with existing benchmarks that focus on evaluating specific fields like network configuration, in this paper, we present **OpsEval**, a comprehensive task-oriented AIOps benchmark designed for LLMs. For the first time, OpsEval assesses LLMs' proficiency in three crucial scenarios (Wired Network Operation, 5G Communication Operation, and Database Operation) at various ability levels (knowledge recall, analytical thinking, and practical application). The benchmark includes 7,200 questions in both multiple-choice and question-answer (QA) formats, available in English and Chinese. With quantitative and qualitative results, we show how various LLM tricks can affect the performance of AIOps, including zero-shot, chain-of-thought, and few-shot in-context learning. We find that GPT4-score is more consistent with experts than widely used Bleu and Rouge, which can be used to replace automatic metrics for large-scale qualitative evaluations.

\*Corresponding author

## CCS CONCEPTS

• **Networks**; • **Software and its engineering** → **Empirical software validation**; • **Information systems** → **Test collections**;

## KEYWORDS

Large language models, AIOps, 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, and Dan Pei. . OpsEval: A Comprehensive Task-Oriented AIOps Benchmark for Large Language Models. In *arXiv preprint*.

## 1 INTRODUCTION

In recent years, large language models (LLMs) have witnessed significant advancements. The latest models, such as GPT-4V [16], GPT-4 [15], LLaMA-2 [6], and ChatGLM2 [3], have demonstrated exceptional generalization capabilities and extensive applicability. As a result, these models have provided numerous opportunities to enhance various downstream domain-specific applications.

Artificial Intelligence for IT Operations (AIOps) plays a crucial role in a number of fields, especially Information Technology (IT) such as cloud computing, and Communication Technology (CT) such as 5G networks. Computer systems and networks are maintained by operations engineers, including anomaly detection, root cause localization, failure mitigation, performance optimization, capacity planning, etc. With its advanced summarizing, report analyzing, and ability of API calls, LLM is well suited for AIOps. Hereinafter, we refer to the LLM used for AIOps as **OpsLLM**. The

challenges of OpsLLM, however, are greater than those of LLM. As with medicine and law, applying LLM to the field of IT or CT operations has a low tolerance for model hallucinations, the phenomenon in which LLMs provide information which is incorrect or inappropriate, presented in a factual manner. It's because the answer of LLM will be directly used to troubleshoot and mitigate failures. LLM's hallucination answers will greatly mislead or even cause bigger failures resulting in economic losses.

Despite the potential benefits of LLM and the performance of OpsLLM, before we get into OpsLLM. The current LLM's performance on AIOps tasks should be clarified. Furthermore, for each change aimed at optimizing LLM to get OpsLLM, such as per-training/fine-tuning, prompt engineering, and model quantitation, a benchmark to provide guidance is essential. Considering the low tolerance for model hallucinations previously mentioned, the current LLM evaluation metric may need to be changed to be suitable for operation evaluation.

C-EVAL [8], AGIEval [27], and MMCU [24] all provide a dataset and platform for general ability assessment. BIG-bench [19] and HELM [10] provide many English tasks and topics for LLMs' general ability evaluation. A number of domain-specific benchmarks are also available, like FinEval [26] and CMB [21], which focus on fields like finance and medicine, evaluate LLMs based on expert knowledge. There aren't too many benchmarks available for OpsLLM. A benchmark for evaluating the performance of LLM in networks is proposed by NetOps [13]. However, we believe that the following criteria should be satisfied for OpsLLM to be a good benchmark.

- The question set must be extracted and processed from reliable sources like textbooks or certification exams without being disclosed to the public, which can avoid the leaderboard being hacked.
- The questions should be categorized into different tasks and abilities to gain more precise comparisons between models.
- While multiple-choice questions are a straightforward way to see how much knowledge LLMs possess, subjective questions and semantic correlated metrics are crucial to evaluate LLMs' ability on practical application.

To address the aforementioned challenges, We introduce **OpsEval**, the first comprehensive Chinese-English bilingual and task-oriented benchmark specifically designed to evaluate **OpsLLM**.

First, we extracted objective questions from various professional books, online sources, and collaborating institutions, which are kept private in OpsEval.

Second, OpsEval includes questions categorized into multi-level abilities (Knowledge Recall, Analytical Thinking, Practical Application) and eight different tasks that are very important in practical applications.

Third, OpsEval comprises both objective questions and expert-crafted subjective questions. At last, OpsEval contains both English and Chinese questions, making it the first OpsLLM benchmark to cover Chinese-English bilingual questions.

The contributions of this paper are as follows:

- We introduce OpsEval, the first comprehensive task-oriented and English-Chinese bilingual AIOps benchmark for large language models, to the best of our knowledge.

- We categorize questions into 8 tasks and 3-level abilities, covering the most common use in practical operations.
- With quantitative and qualitative results, we show how various LLM tricks can affect the performance of AIOps, including chain-of-thought and few-shot in-context learning.
- We quantitatively show that GPT4-score is more consistent with experts than widely used Bleu and Rouge, which can be used to replace automatic metrics for large-scale qualitative evaluations.

This paper is organized as follows. We first introduce related works on benchmarks for LLMs in Sec. 2. Then we provide a detailed introduction of OpsEval in Sec. 3. In Sec. 4, we elaborate the experiment design to evaluate LLMs in OpsEval. In Sec. 5, we present the results and observations of evaluations on various LLMs. Finally, we draw our conclusion in Sec. 6.

## 2 RELATED WORKS

As LLMs continue to evolve rapidly, their complex and varied capabilities are becoming increasingly recognized. Traditional NLP evaluation metrics, however, fall short of accurately gauging these abilities. As a result, there's a growing trend towards proposing evaluation benchmarks tailored specifically for LLMs. These can generally be divided into two categories: general ability benchmarks and domain-specific benchmarks.

General ability benchmarks serve to assess the general abilities of LLMs across a variety of tasks. These tasks test for common sense, general knowledge, reasoning ability and so on, rather than being limited to a specific domain. HELM [10] employs 7 distinct metrics in 42 unique scenarios, offering a comprehensive evaluation of LLMs' capabilities across multiple dimensions. BIG-bench [19] comprises 204 tasks that span a wide array of topics, with a particular focus on tasks deemed beyond the reach of current LLMs. C-Eval [8] is the first comprehensive Chinese evaluation suite, designed to rigorously assess the advanced knowledge and reasoning abilities of Chinese LLMs, as well as their understanding of context-specific knowledge unique to the Chinese language. AGIEval [27] curates authentic questions from examinations such as the Chinese College Entrance Exam and the SAT, thereby constructing an evaluation dataset that is fundamentally human-centric. In a parallel endeavor, MMCU [24] leverages questions sourced from the Chinese College Entrance Exam and various professional qualification examinations to establish a robust benchmark specifically designed to assess the capabilities in understanding. CG-Eval [25] focuses on assessing the generation capabilities of LLMs, employing a testing framework that includes term definitions, short-answer questions, and computational problems.

Domain-specific benchmarks evaluate the abilities of LLMs to handle tasks in specific fields. These benchmarks generally require LLMs to possess specialized knowledge in a certain 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 significantly less attention.

FinEval [26] is a benchmark designed specifically to measure the advanced financial knowledge of Chinese LLMs. MultiMedQA [18]

is an extensive medical question-and-answer dataset, with questions derived from professional medical exams, research, and consultation records. It demands a deep understanding of medical knowledge and includes both multiple-choice and open-ended questions. Huatuo-26M [9] and CMB [21] are comprehensive medical question-and-answer datasets. Huatuo-26M [9] comprises a substantial number of actual medical consultation records and some medical knowledge question-and-answer content. CMB [21] includes multiple-choice qualification examination questions (CMB-Exam) and complex clinical diagnostic questions based on real case studies (CMB-Clin), with the correct answers established through expert consensus.

NetOps [13] focuses on evaluations in the network field, which is relevant to the field of Ops. NetOps includes multiple-choice questions in both English and Chinese, as well as a small number of filling-blanks and question-answering questions.

In contrast, our research encompasses a diverse range of Ops sub-domains, including Wired Network Operation, 5G Communication Operation, and Database Operation. Furthermore, we have meticulously delineated the categorization of tasks and abilities, thereby providing a more nuanced and comprehensive assessment framework. We have also adopted a wider variety of more reasonable metrics for the evaluation of subjective questions, and conducted a detailed analysis of the performance of LLM in various tasks and abilities. Furthermore, we explore the influence of diverse quantization parameters of LLMs on the performance of handling tasks in AIOps.

### 3 OPSEVAL BENCHMARK

In our evaluation of LLMs within the AIOps domain, we categorize our assessment questions into objective and subjective questions.

Objective questions, usually framed as multiple-choice questions, offer a structured approach with definitive answers. These questions are straightforward and provide a clear metric for assessment.

However, given the intricacies of advanced models, they may be influenced by the options provided, which can lead to their responses being driven more by pattern recognition rather than a proper understanding of the content.

Subjective questions do not come with predefined options. This necessitates the model to rely more on its comprehension and knowledge base, offering a clearer insight into its cognitive capabilities. Such questions can better assess LLMs' ability to generate coherent and contextually relevant responses.

By incorporating both questions, we aim for a comprehensive and balanced assessment. This ensures that the evaluated models are not merely recognizing patterns but also showcasing genuine comprehension of diverse tasks.

For fairness of the evaluation, we keep our questions private and not disclosed. To evaluate a new model, users can submit a Docker image with an initialization script that takes a prompt as input and outputs the model's answer when starting a container based on it. We will run the evaluation automatically and obtain the result on the website of OpsEval. Users can choose to disclose their results on the leaderboard of OpsEval or not based on their preference.

#### 3.1 Objective Questions

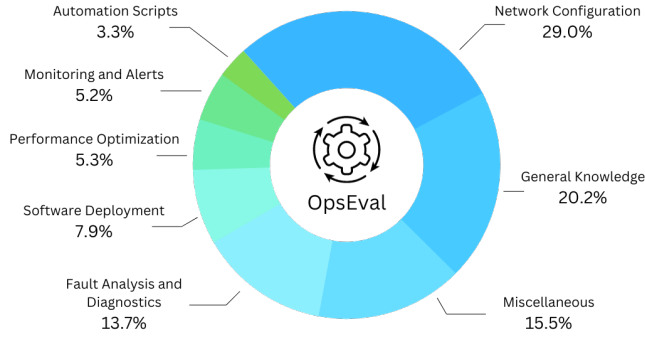
**Data source:** The objective questions have been sourced from exams related to well-known international certifications in the AIOps domain, which are held in high esteem globally. Our collection process involves gathering questions from various books, online resources, and collaborating institutions. These questions are primarily formatted as multiple-choice, encompassing single-answer and multiple-answer styles. Each question has a prompt, a possible answer choice, and a relevant explanation or analysis. The main scenarios we focused on include Network Operation, Communication Technology Operation, and Database Operation. We will also continue to refine and expand the assessment scenarios in the future.

**Data Processing:** The processing of our original test set was systematically carried out in several stages:

- (1) *Initial Screening:* In this phase, we aim to filter out questions that are not directly related to AIOps. To assist in this process, we employ GPT-3.5, leveraging its advanced capabilities to determine the relevance of each question to AIOps.
- (2) *Duplication Removal:* Any repeated or highly similar questions are identified and removed to avoid redundancy in the test set.
- (3) *Content Filtering:* We exclude questions that predominantly depend on the context of articles or external content.
- (4) *Format Standardization:* We simplify and standardize the format of each question. The adopted format is defined as a tuple  $(Q, A)$ , where  $Q$  represents the question with its options, and  $A$  includes the correct answers along with their analyses.
- (5) *Manual Review:* A subsequent round of manual screening is performed to ensure the quality and relevance of the test set. This meticulous process resulted in a refined test set of approximately 7,000 objective questions, among which 55.7% are related to Wired Network Operation, 37.1% are related to 5G Communication Operation, and 7.2% are related to Database Operation.

**Task Categorization:** In the complex landscape of operations and maintenance, recognizing the multidimensional nature of tasks and challenges is essential. For a comprehensive evaluation of LLMs within the AIOps domain, we devise a categorization that captures many tasks that professionals confront in practical applications. The formulation of our eight distinct operation and maintenance tasks is influenced by industry relevance, task frequency, and the significance of each area in AIOps. Details of the eight scenarios can be found in Appendix A.1. The distribution of the objective questions across these eight categories is depicted in Figure 1.

**Ability Categorization:** A comprehensive evaluation of a model goes beyond just assessing its ability to produce correct answers. It is equally crucial to understand the depth, complexity, and applicability of its reasoning across different levels of cognitive demands. With this perspective, based on which ability is required to answer them, we manually classify all questions into three categories. The three abilities are Knowledge Recall, Analytical thinking, and Practical Application, reflecting the challenges professionals might



**Figure 1: Scenario Categorization in OpsEval**

encounter in real-world scenarios. Details of the three abilities can be found in Appendix A.2.

### 3.2 Subjective Questions

**Data Collection:** The subjective questions within the OpsEval test set are sourced from a blend of carefully curated resources to ensure comprehensiveness and relevance:

- *Generated from Objective Questions:* A portion of our subjective questions is derived from carefully selected objective questions from our original test set. These questions are transformed into a subjective format after being identified for their potential depth and breadth.
- *Extracted from Books:* To enhance the diversity and depth of our test set, we also source subjective questions from authoritative books covering a range of AIOps domains. This ensures that our test set is extensive and aligns with industry standards and current best practices.

**Data Processing:** Upon collection, a rigorous process is followed to refine and standardize the subjective questions:

- (1) *Question Summarization:* The objective questions that were transformed into subjective questions underwent a summarization process. This involves distilling the essence of each question and presenting it in an open-ended format without predefined options.
- (2) *Inclusion of Reference Texts:* For questions generated with the assistance of GPT-4, reference texts are provided in the prompt to guide the generation process and ensure accuracy.
- (3) *Data Structuring:* Each subjective question is meticulously structured to include the raw question, critical points of the answer, a detailed answer, the task, and the associated ability. This structured approach facilitates easy evaluation and analysis.

In total, we accumulate a collection of **200** subjective questions, ensuring a well-rounded evaluation of model capabilities in the AIOps domain. An illustrative example of a saved question can be found in Appendix A.3.

### 3.3 Evaluation Metrics

We use different metrics for objective and subjective questions (i.e., multiple-choice questions).

For objective questions, we use **accuracy** as the metric. As LLMs may output more content than the options, the answers of LLMs are extracted from their raw replies by an Option Extractor based on regular expressions. Based on the extracted answer and the ground-truth labels, we calculate the accuracy. The baseline accuracy of objective questions is lower than 25% as we have multiple-answer, multiple-choice questions among all questions.

For subjective questions, we use two types of metrics, where one is based on word overlaps and the other is based on semantic similarity. For the first type, we use Rouge [11] and Bleu [17], widely used in NLP tasks, especially in the translation task. For the second type, we use GPT-4 and experts to obtain the output score of LLMs, called GPT4-Score and Expert-Evaluation, specifically in OpsEval.

**Rouge** (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics

for evaluating automatic summarization and machine translation.

ROUGE-N is the overlap of n-grams between the system and reference summaries. ROUGE-L considers sentence-level structure similarity naturally and automatically identifies the longest co-occurring in sequence n-grams. Briefly speaking, Rouge can be understood as the recall of the ground-truth answer. We utilize the *rouge\_score* python package to calculate Rouge1, Rouge2, and RougeL in OpsEval. The score of Rouge is normalized from 0 to 100. The higher the score is, the better it is.

**Bleu** (Bilingual Evaluation Understudy) can be understood as the precision of the generated answer. We utilize the *scorebleu* python package to calculate Bleu in OpsEval. The score of Bleu is normalized from 0 to 100. The higher the score is, the better it is.

**GPT4-Score** is a score generated by GPT4 with a deliberately crafted prompt. Scoring by LLMs is used increasingly, especially after the parameters of LLMs get larger. We compose the scoring prompt of the question, the ground-truth keypoint, the ground-truth detailed answer, and the answer of LLM to be scored. The score is between 1 and 10, and the higher is better.

**Expert-Evaluation** is designed by us for manually scoring LLMs' outputs based on three criteria highly related to operations' needs. The three criteria in consideration are as follows:

- **Fluency.** Assessment of the linguistic fluency in the model's output, compliance with the subjective question's answering requirements, and the presence or absence of paragraph repetitions or irrelevant text.
- **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.

For each output to a question of an LLM, we asked experts to score it between 0 and 3 for each criterion. During the scoring, the raw question, key points of the answer, the detailed answer, and the output of an anonymous model are given at each iteration. Since avoiding bias stemming from preconceived notions about existing

models is essential, no information about the anonymous model is given.

## 4 EXPERIMENT DESIGN

In this section, we will show the experiment design of OpsEval. We evaluate various LLMs on OpsEval, aiming to understand multiple abilities of different LLMs in addressing different question types (objective and subjective questions) and tasks. We also evaluate LLMs with different quantization parameters on OpsEval.

### 4.1 Models

**Table 1: Models evaluated in this paper**

Model	Creator	#Parameters	Access <sup>1</sup>
GPT-4	OpenAI	<i>undisclosed</i>	API
ChatGPT	OpenAI	<i>undisclosed</i>	API
LLaMA-2-70B	Meta	70B	Weights
LLaMA-2-13B	Meta	13B	Weights
LLaMA-2-7B	Meta	7B	Weights
Chinese-LLaMA-2-13B	Cui et al.	13B	Weights
Chinese-Alpaca-2-13B	Cui et al.	13B	Weights
Baichuan-13B-Chat	Baichuan Intelligence	13B	Weights
ChatGLM2-6B	Tsinghua	6B	Weights
InternLM-7B	Shanghai AI Laboratory	7B	Weights
Qwen-7B-Chat	Alibaba Cloud	7B	Weights

**Table 2: GPTQ models for LLaMA-2-70B**

Model	Size	#GPTQ Dataset	Disc
LLaMA-2-70B	140GB	/	Raw LLaMA-2-70B model.
LLaMA-2-70B-Int4	35.33GB	wikitext	4-bit quantization model.
LLaMA-2-70B-Int3	26.78GB	wikitext	3-bit quantization model.

As shown in Table 1, we evaluate the popular LLMs covering different weights and organizations that can process both English and Chinese input. The detailed information of all LLMs in Table 1 can be found in Appendix B.1.

Besides, we evaluate LLaMA-2-70B with multiple quantization parameters to get an overview of the effect of different quantization parameters. Specifically, we use GPTQ [7] models with 3-bit and 4-bit quantization parameters.<sup>2</sup> The size of LLaMA-2-70B is calculated based on 70B parameters. The two GPTQ models evaluated in our experiments are calibrated on *wikitext* [12], a language modeling dataset extracted from the set of verified Good and Featured articles on Wikipedia. GPTQ is a post-training quantization (PTQ) method to make the model smaller with a calibration dataset. It is a one-shot weight quantization method based on approximate, highly accurate, and efficient second-order information. The details of GPTQ models can be found in Table 2.

<sup>1</sup>The "access" column in the table shows whether we have full access to the model weights or can only access them through API.

<sup>2</sup>The two quantization models are downloaded from <https://huggingface.co/TheBloke/Llama-2-70B-chat-GPTQ>.

## 4.2 Setups

**4.2.1 Objective Questions.** To get a comprehensive overview of the performance of popular LLMs on OpsEval, we use as many settings as possible to perform the evaluation. We evaluate LLMs in zero and few-shot settings (3-shot in our implementation). For zero-shot settings, we want to evaluate the abilities of LLMs without any examples from a user’s perspective, as users will not provide examples in ordinary use. For few-shot settings, we aim to assess the potential of the LLMs from a developer’s perspective, which can obtain a better performance than zero-shot settings. For each setting, we evaluate LLMs in four sub-settings of prompt engineering, that is, naive answers (Naive), self-consistency (SC [22]), chain-of-thought (CoT [23]), self-consistency with chain-of-thought (CoT+SC). As we have English and Chinese questions, we design prompts for the two languages.

**Naive.** The Naive setting is to expect the LLMs to generate the answer without any other explanations. Since we have the task type of each question, we integrate the task into the prompt.

**SC.** Self-consistency is selecting the most consistent answer among several queries on LLMs. Although it aims "to replace the naive greedy decoding used in chain-of-thought prompting," it can generate naive answers as it may generate different answers with the same prompts. We set the number of queries in SC to 5.

**CoT.** The CoT setting aims to enable LLMs to obtain complex reasoning capabilities through intermediate reasoning steps. We construct specific prompts for CoT setting in both zero and few-shot evaluations. Details of the prompt construction can be found in B.1.

**CoT+SC.** We combine CoT and SC to boost the performance of CoT prompting. By SC, We choose the consistent reasoning path and answer several of the same queries. Like the SC setting, we set the number of queries in CoT+SC to 5.

**4.2.2 Subjective Questions.** We combine the scenario and ability of each question and the question as the prompt for LLMs. As we want to simulate the daily use of LLMs as a typical user and expect LLMs to generate answers according to the question, we do the zero-shot evaluation on LLMs in the Naive setting. An example of constructing the prompt can be found in Figure 10 in Appendix A.3.

## 5 EVALUATION

### 5.1 Overall Performance

The results of zero-shot and few-shot evaluation with four settings on the Wired Network Operation test set are shown in Table 3. Results on the other two test sets are shown in Appendix B.3. Performance on both English and Chinese questions are presented in the tables.<sup>3</sup> From the overall performance results, we can come to several findings.

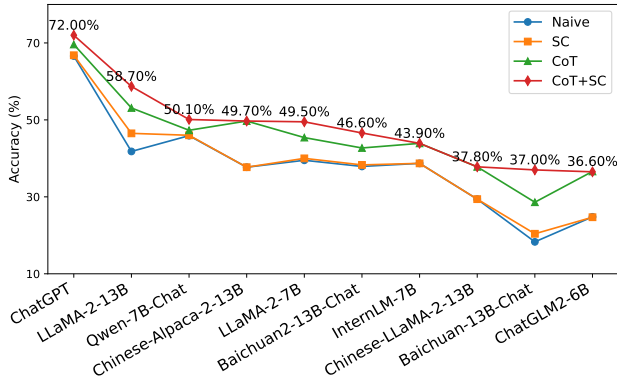
On both English and Chinese questions, GPT-4 consistently outperforms all other models, surpassing the best performances of all other LLMs. In the Wired Network Operation test set, LLaMA-2-13B and Baichuan-13B-Chat, *when employing the Self-Consistency and*

<sup>3</sup>Due to the consideration of time, cost, and API rate limits, for GPT-4 we only make the 3-shot evaluation with the CoT setting as the upper bound of all LLMs to provide a reference.

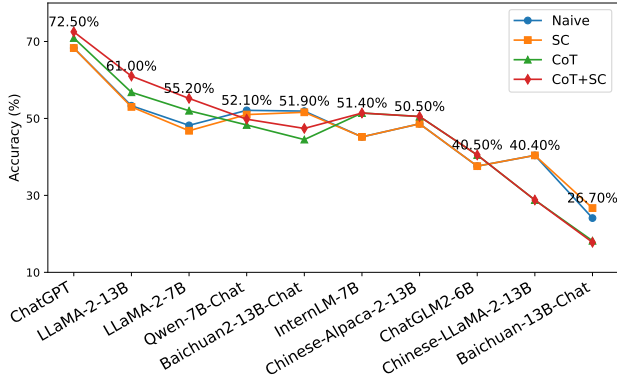
**Table 3: LLMs’ overall performance on Wired Network Operation test set**

Model	English Test Set								Chinese Test Set							
	Zero-shot				3-shot				Zero-shot				3-shot			
	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC
GPT-4	/	/	/	/	/	/	<b>88.70</b>	/	/	/	/	/	/	/	<b>86.00</b>	/
ChatGPT	<u>66.60</u>	<u>66.80</u>	<u>69.60</u>	<u>72.00</u>	<u>68.30</u>	<u>68.30</u>	70.90	<b>72.50</b>	<u>58.40</u>	<u>58.60</u>	<u>64.80</u>	<b>67.60</b>	<u>59.20</u>	<u>59.70</u>	65.20	<u>67.40</u>
LLaMA-2-13B	41.80	46.50	53.10	58.70	53.30	53.00	56.80	<b>61.00</b>	29.70	31.60	51.60	<b>57.00</b>	39.60	38.90	48.00	50.60
Baichuan2-13B-Chat	37.90	38.30	42.70	46.60	<b>51.90</b>	51.60	44.50	47.45	44.60	45.40	41.60	44.30	45.60	45.70	43.90	<b>46.70</b>
Baichuan-13B-Chat	18.30	20.40	28.60	<b>37.00</b>	24.10	26.70	18.20	17.80	15.20	16.00	43.90	49.70	34.30	36.10	51.30	<b>55.60</b>
Chinese-Alpaca-2-13B	37.70	37.70	49.70	49.70	48.60	48.60	<b>50.50</b>	<b>50.50</b>	33.10	33.10	<b>44.20</b>	<b>44.20</b>	44.00	44.00	42.70	42.70
Chinese-LLaMA-2-13B	29.40	29.40	37.80	37.80	<b>40.40</b>	<b>40.40</b>	28.80	28.80	22.50	22.50	38.80	38.80	<b>41.80</b>	41.80	32.20	32.20
LLaMA-2-7B	39.50	40.00	45.40	49.50	48.20	46.80	52.00	<b>55.20</b>	29.80	30.20	50.10	<b>55.60</b>	38.60	40.80	45.60	50.40
Qwen-7B-Chat	45.90	46.00	47.30	50.10	<b>52.10</b>	51.00	48.30	49.80	29.60	29.90	50.60	<b>53.50</b>	50.40	46.90	46.90	47.70
InternLM-7B	38.70	38.70	43.90	43.90	45.20	45.20	<b>51.40</b>	<b>51.40</b>	41.70	41.70	38.40	38.40	<b>42.60</b>	<b>42.60</b>	41.30	41.30
ChatGLM2-6B	24.80	24.70	36.60	36.50	37.60	37.60	<b>40.50</b>	40.50	33.80	33.70	42.10	<b>42.20</b>	36.00	36.00	39.50	39.50

Note: The best accuracy of all settings for each LLM is in **bold** font. The best accuracy of all LLMs for each setting is underlined.



**Figure 2: Accuracy on English Wired Network Operation test set (zero-shot)**



**Figure 3: Accuracy on English Wired Network Operation test set (3-shot)**

Chain of Thought (CoT) prompt methods, approach the performance of ChatGPT in both English and Chinese test sets, respectively.

Smaller models, such as LLaMA-2-7B and InternLM-7B, exhibit competitive performance in objective questions, approaching the capabilities of models with 13B parameters, giving credits to their fine-tuning process and the quality of their training data.

Furthermore, the effectiveness of the four prompt settings varies across different LLMs. LLMs fine-tuned specifically for Chinese exhibit better performance on English and Chinese test sets compared to LLMs that have not undergone Chinese fine-tuning. We discuss further insights into these observations in Sec. 5.6.

## 5.2 Performance Under Different Settings

For both English and Chinese test sets, we examine LLMs’ zero-shot and few-shot performances under four settings mentioned earlier in Sec. 4. Results are shown in Figure 2 and Figure 3. Based on Figure 2(as well as the results of other test sets), we can conclude the following observations:

- (1) For most models, the performance improves from the setting of Naive to SC, CoT, and SC+CoT. *Notably, few-shot performance is better than zero-shot performance.*
- (2) Among these settings, CoT prompts yield the most significant improvement in LLMs’ answering capability. *SC prompts result in relatively minor improvements, as LLMs’ responses tend to be consistent across repeated questions, aligning with the desired outcome in operational tasks where reliability and consistency are essential.*
- (3) In few cases, more advanced evaluation methods surprisingly lead to poorer results. The detailed analysis for this can be found in Appendix B.5.1.

## 5.3 Performance on Different Tasks and Abilities

To investigate how LLMs perform in each operation task and to what extent they possess the abilities of Knowledge Recall, Analytical Thinking, and Practical Application, we summarize the result of different parameter-size LLM groups based on the task and ability classification mentioned in Sec. 3.1 and plot them on two radar charts, Figure 6, concerning their ability performance and task performance respectively.

In terms of the eight tasks we tested, LLMs generally yield higher accuracy in General Knowledge tasks, while their performance drop and vary 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.

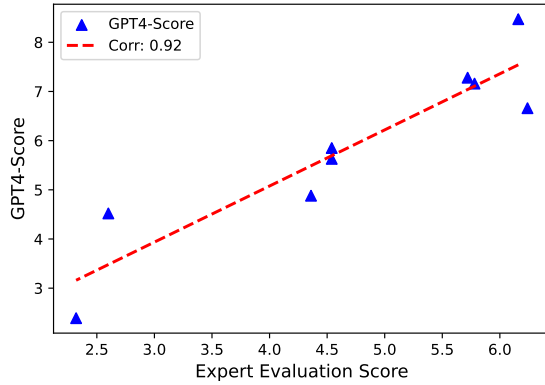
Among the three abilities, LLMs perform best in Practical Application, followed by Knowledge Recall. It is expected that LLMs



perform poorly in Analytical Thinking questions, as accurately deducing conclusions from existing facts for LLMs remains a challenging research topic. LLMs perform best in Practical Application because LLMs we test are trained on the corpus where best operation practices are involved, familiarizing the LLMs with solutions to many real-world tasks.

By grouping LLMs by their parameter size, we find that although LLMs with 13B parameters have higher accuracy in their best cases compared with LLMs with no more than 7B parameters, *different 13B LLMs' performance varies drastically*, causing its lower bound to be even lower than that of 7B. *LLMs with no more than 7B parameters, on the other hand, have a more stable performance range within the group.*

#### 5.4 Performance on Subjective Questions



**Figure 4: Scatter plot and trendline of different automated metric scores compared to Expert Evaluation scores.**

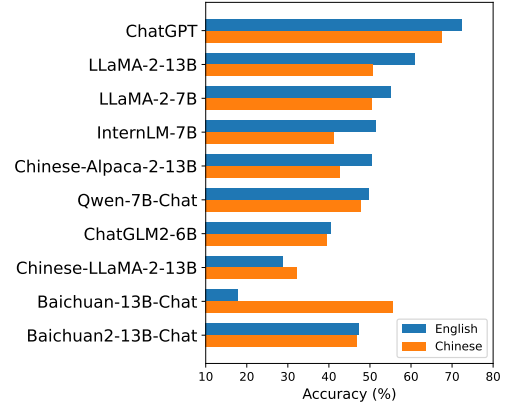
Table 4 presents the evaluation results of 200 subjective English questions across four metrics: Rouge, Bleu, GPT4-Score, and Expert-Evaluation, sorted by GPT4-Score results.

The rankings based on Rouge and Bleu scores do not align well with GPT4-Score and Expert-Evaluation, as shown in table 5. LLMs with poor actual performance may generate keywords, resulting in higher Rouge and Bleu scores. In contrast, LLMs with good performance might receive lower Rouge/Bleu scores due to differences in wording compared to the standard answers.

Regarding to GPT4-Score, the rankings closely resemble those based on Expert-Evaluation. In table 6, we calculate the correlation coefficients between GPT4-Score and different sub-metrics of Expert-Evaluation to gain more insights. Among the three metrics, *rankings of GPT4-Score align most closely with the Accuracy metric, suggesting that GPT4 is most reliable on the factuality with its vast knowledge base.* The format and length of the generated content also heavily influence GPT4-Score, as suggested by the high positive correlation between GPT4-Score and Fluency. On the other hand, there are more discrepancies in rankings concerning the Evidence metric, indicating that GPT4-Score needs to fully consider the role of arguments and evidence in cases where answers are ambiguous.

In the Expert-Evaluation, where Evidence is a significant criterion, LLMs with more elaborate arguments, such as Chinese-Alpaca-2-13B, can outperform ChatGPT in total scores even when their Accuracy scores are much lower than the latter.

Here, we calculate the Total score in Expert-Evaluation by assigning equal weights to Fluency, Accuracy, and Evidence, which should be only considered as one of the many ways to value those aspects. Different weights should be allocated to sub-metrics of Expert-Evaluation in specific applications based on the real-world scenarios.



**Figure 5: LLMs' few-shot performance on English/Chinese test set (CoT+SC)**

#### 5.5 Performance on Different Quantization parameters

Figure 7 shows the accuracy of LLaMA-2-70B of different quantization parameters on English objective questions. We do both few-shot and zero-shot evaluations with the naive setting. Using quantization during inference brings a performance degradation of an LLM.

LLaMA2-70B-Int4 can achieve an accuracy close to LLaMA-2-70B without quantization. Specifically, on English objective 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. As 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.

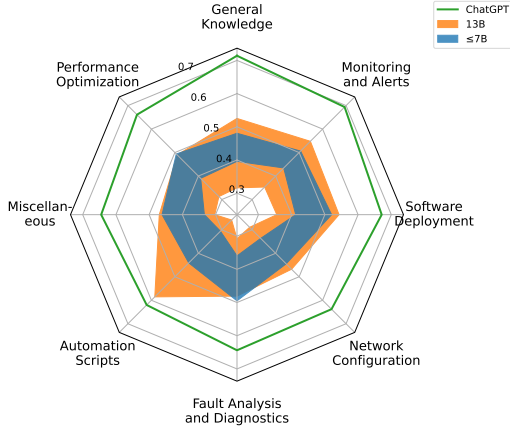
However, *LLaMA2-70B-Int3 has a performance degradation that cannot be ignored*, as shown in Figure 7. On average, the accuracy of LLaMA2-70B-Int3 has a 12.46% degradation compared to LLaMA-2-70B and a 9.30% degradation compared to LLaMA2-70B-Int4. The reason may be that the information of the full-sized model is lost too much in 3-bit quantization.

#### 5.6 Performance on Different Languages

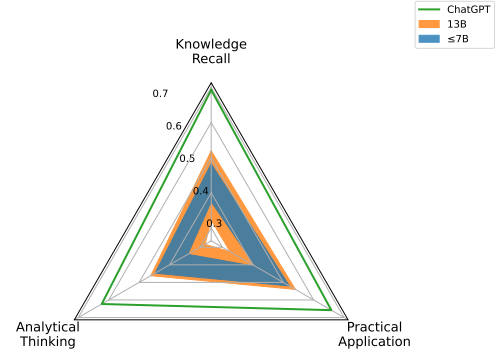
In Figure 5, we compare the few-shot performance of various LLMs under the CoT+SC setting for both English and Chinese questions. Notably, some of the LLMs that have undergone specific training or fine-tuning with Chinese language corpus, such

**Table 4: LLMs’ performance on English network operation subjective problems**

Model	Rouge			Bleu Score	GPT4-Score	Expert Evaluation			
	Rouge1	Rouge2	RougeLsum			Fluency	Accuracy	Evidence	Total
ChatGPT	13.38	5.65	12.26	6.78	8.47	3.00	1.96	1.20	6.16
LLaMA2-70B-Int4	8.69	2.51	7.74	4.20	7.28	2.92	1.48	1.32	5.72
LLaMA2-13B-Chat	5.75	1.68	4.98	3.43	7.16	2.82	1.34	1.62	5.78
Chinese-Alpaca-2-13B	3.48	0.96	3.25	1.85	6.66	2.90	1.52	1.82	6.24
Baichuan-13B-Chat	5.58	1.85	4.76	0.35	5.85	2.40	1.12	1.02	4.54
Qwen-7B-Chat	13.03	4.76	11.82	4.33	5.63	2.56	1.14	0.84	4.54
ChatGLM2-6B	10.43	3.24	9.71	5.07	4.88	2.84	0.76	0.76	4.36
InternLM-7B-Chat	14.34	5.39	13.27	0.54	4.52	1.80	0.70	0.10	2.60
Chinese-LLaMA-2-13B	9.18	2.90	9.19	0.24	2.39	1.42	0.72	0.18	2.32



**(a) Performance on 8 tasks**



**(b) Performance on 3 abilities**

**Figure 6: LLMs’ performance on different tasks and abilities.** LLMs’ results are grouped by their parameter size: "<7B" includes Qwen-7B-Chat, ChatGLM2-6B, InternLM-7B and LLaMA-2-7B; "13B" includes Baichuan-13B-Chat, LLaMA-2-13B, Chinese-Alpaca-2-13B, Chinese-LLaMA-2-13B. We calculate the lower and upper bound of LLMs’ performances to give a rough range on each task/ability.

**Table 5: Correlation coefficients between Expert-Evaluation Total and other metrics**

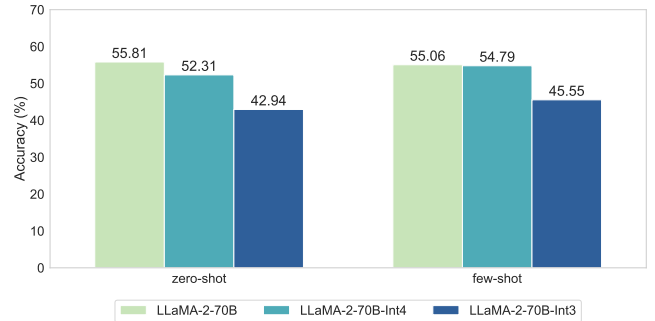
Metric	GPT4-Score	Bleu-Score	RougeLsum
Correlation coefficient	0.9211	0.6108	-0.4559

**Table 6: Correlation coefficients between GPT4-Score and sub-metrics of Expert-Evaluation**

Metric	Fluency	Accuracy	Evidence	Total
Correlation coefficient	0.8700	0.9084	0.7978	0.9211

as Chinese-Alpaca-2-13B, Qwen-7B-Chat, and ChatGLM2-6B, still perform better in answering English questions than Chinese ones. We analyze potential reasons for this phenomenon as follows:

- (1) Some LLMs may have been trained using original English books from the sources of our test set, which could give them a natural advantage in answering English questions.
- (2) During the translation process from English to Chinese, there may be deviations in the context of certain questions



**Figure 7: LLaMA-2-70B’s performance of different quantization parameters.** Both zero-shot and few-shot evaluations have been conducted on English Wired Network Operation test set using the naive setting.

due to limitations in translation software and human review, potentially affecting the quality of the Chinese questions.

- (3) LLMs that have been fine-tuned specifically for the Chinese language may exhibit improved Chinese answering abilities but might experience a decrease in their understanding of



question formats and the comprehension of prompts related to CoT.

Despite the observed fact that performance tends to be lower for Chinese questions compared to the original English questions, we can still glean valuable insights into the language capabilities of the LLMs. Notably:

- (1) ChatGLM2-6B experiences the smallest decline in performance when transitioning to Chinese questions. *This improvement can be attributed to its substantial exposure to Chinese language data during training rather than simple fine-tuning on top of an existing base model.*
- (2) LLaMA-2-13B exhibits the most significant drop in performance when switching to Chinese questions. *This indicates that the shift in language impacts LLMs' general understanding ability and capacity to extract domain-specific knowledge.*

We also observe an interesting phenomenon with the Baichuan-13B-Chat in the 3-shot evaluation with the CoT+SC setting, where its performance in Chinese questions significantly outperforms in English. We examine the LLM's outputs and analyze a sample question to shed light on this phenomenon in Appendix B.5.2.

## 6 CONCLUSION

In this paper, we introduce **OpsEval**, a comprehensive task-oriented AIOps benchmark designed for LLMs. Unprecedented in its approach, OpsEval evaluates LLMs' proficiency across three pivotal scenarios (Wired Network Operation, 5G Communication Operation, and Database Operation) while considering varying ability levels encompassing knowledge recall, analytical thinking, and practical application. This comprehensive benchmark comprises 7,200 questions in both multiple-choice and question-answer (QA) formats, presented in both English and Chinese.

Supported by quantitative and qualitative results, elucidates the nuanced impact of various LLM techniques, such as zero-shot, chain-of-thought, and few-shot in-context learning, on AIOps performance. Notably, the GPT4-score emerges as a more reliable metric when compared to widely used Bleu and Rouge, suggesting its potential as a replacement for automatic metrics in large-scale qualitative evaluations.

The identified flexibility within the OpsEval framework presents opportunities for future exploration. The adaptability of this benchmark facilitates the seamless integration of additional fine-grained tasks, providing a foundation for continued research and optimization of LLMs tailored for AIOps.

## REFERENCES

- [1] 2023. baichuan-inc/Baichuan-13B. <https://github.com/baichuan-inc/Baichuan-13B>
- [2] 2023. QwenLM/Qwen-7B. <https://github.com/QwenLM/Qwen-7B>
- [3] 2023. THUDM/ChatGLM2-6B. <https://github.com/THUDM/ChatGLM2-6B>
- [4] Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca. *arXiv preprint arXiv:2304.08177* (2023). <https://arxiv.org/abs/2304.08177>
- [5] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 320–335.
- [6] Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288* [cs.CL]
- [7] Elias Frantar, Saleh Ashkboos, Torsten Hoeftler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323* (2022).
- [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] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26M, a Large-scale Chinese Medical QA Dataset. *arXiv e-prints* (2023), arXiv–2305.
- [10] 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.
- [11] 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>
- [12] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv:1609.07843* [cs.CL]
- [13] Yukai Miao, Yu Bai, Li Chen, Dan Li, Haifeng Sun, Xizheng Wang, Ziqiu Luo, Dapeng Sun, Xiuting Xu, Qi Zhang, Chao Xiang, and Xinchu 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. 2022. ChatGPT: Optimizing Language Models for Dialogue. *OpenAI Blog* (2022). <https://openai.com/blog/chatgpt/>
- [15] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [16] OpenAI. 2023. GPT-4V(ision) System Card. [https://cdn.openai.com/papers/GPTV\\_System\\_Card.pdf](https://cdn.openai.com/papers/GPTV_System_Card.pdf)
- [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. <https://doi.org/10.3115/1073083.1073135>
- [18] 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).
- [19] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shueb, 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.
- [20] InternLM Team. 2023. InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities. <https://github.com/InternLM/InternLM>.
- [21] 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.
- [22] 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]
- [23] 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]
- [24] Hui Zeng. 2023. Measuring Massive Multitask Chinese Understanding. *arXiv e-prints* (2023), arXiv–2304.
- [25] Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2023. Evaluating the Generation Capabilities of Large Chinese Language Models. *arXiv e-prints* (2023), arXiv–2308.
- [26] Liwen Zhang, Weige Cai, Zhaoawei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. *arXiv e-prints* (2023), arXiv–2308.
- [27] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv e-prints* (2023), arXiv–2304.

## A DETAILS OF OPSEVAL BENCHMARK

### A.1 Task Types of Questions

We categorize all questions in OpsEval into 8 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 dissemination and management of software applications throughout the network or system, verifying their correct installation.
- *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.

### A.2 Ability Levels of Questions

Different questions require different levels of ability to answer. We classify all questions in OpsEval into 3 categories. The details of each ability are as follows:

- (1) *Knowledge Recall*: Questions under this category primarily test a model’s capacity to recognize and recall core concepts and foundational knowledge. Such questions are akin to situations where a professional might need to identify a standard procedure or recognize a well-known issue based solely on previous knowledge.
- (2) *Analytical thinking*: These questions demand more than mere recall. They necessitate a deeper level of thought, expecting the model to dissect a problem, correlate diverse pieces of information, and derive a coherent conclusion. It mirrors real-world scenarios where professionals troubleshoot complex issues by connecting various dots and leveraging their comprehensive understanding.
- (3) *Practical Application*: These questions challenge a model’s ability to apply its foundational knowledge or analytical conclusions to provide actionable recommendations for specific scenarios. It epitomizes situations where professionals are expected to make decisions or suggest solutions based on in-depth analysis and expertise.

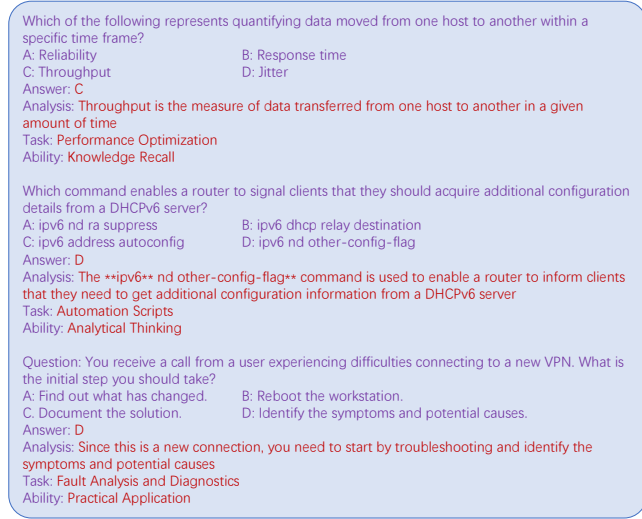


Figure 8: Three examples of the processed questions

Figure 8 illustrates examples in our question set, shedding light on our classification methodology.

### A.3 An Example of Subjective Questions

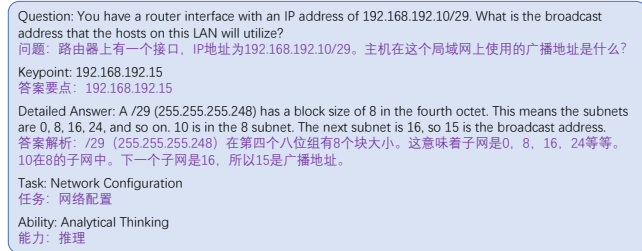


Figure 9: An example of the saved subjective questions in OpsEval

A saved subjective question in OpsEval is presented in Figure 9, which contains not only the raw question but also its type of task.

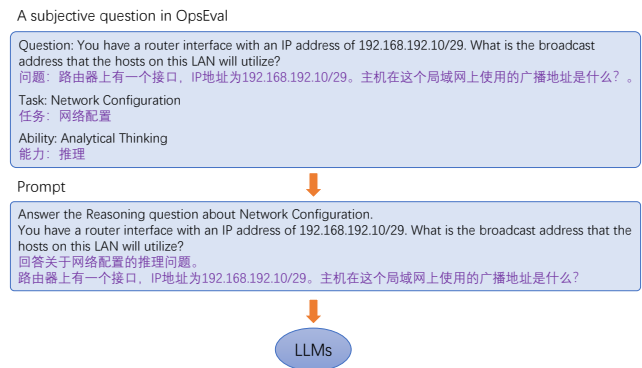


Figure 10: An example of building the prompt of subjective questions.

As shown in Figure 10, we combine the task and ability of each question with the question itself as the prompt for LLMs.

## B ADDITIONAL DETAILS OF EXPERIMENTS

### B.1 Detailed Information of LLMs Evaluated

GPT-4 [15] is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks. It is recognized as the strongest language model currently. ChatGPT [14] is an earlier AI-powered language model developed by OpenAI which is built upon GPT-3.5. We use the gpt-3.5-turbo version in our experiments. Llama 2 [6] is a second-generation open-source LLM from Meta which is very popular due to its open-source feature. It has the ability to process multiple languages including Chinese. We evaluate three weights (70B, 13B and 7B as shown in 1) of Llama 2. Although Llama 2 is able to process Chinese input, it has a small Chinese vocabulary so that its ability of understanding and generating Chinese text is limited. As a result, Many new LLMs is derived from Llama 2 to improve its ability on Chinese text. Chinese-LLaMA-2 [4] expands and optimizes the Chinese vocabulary based on the original Llama 2, using a large scale of Chinese texts for continue pretraining, further enhancing the abilities of basic semantic understanding on Chinese. Chinese-Alpaca-2 [4] a chat model which is instruction-tuned based on Chinese-LLaMA-2. Furthermore, we evaluate some Chinese-oriented LLMs which are published by institutions in China. Baichuan-13B-Chat [1] is aligned chat model based on Baichuan-13B-Base [1] which is an open-source LLM published by Baichuan Intelligence. ChatGLM2 [3] is a second-generation chat model from Tsinghua Knowledge Engineering Group, based on General Language Model architecture (GLM [5]) trained on English and Chinese data. InternLM [20], developed by Shanghai AI Laboratory, is a multi-lingual model based on billions of parameters through multi-stage progressive training on over trillions of tokens. Qwen [2] (abbr. Tongyi Qianwen) is a series of LLMs developed by Alibaba Cloud. And Qwen-7B-Chat, a large-model-based AI assistant, which is trained with alignment techniques based on the pretrained Qwen-7B.

### B.2 An Example of CoT Prompt

For zero-shot evaluation in the CoT setting, we get the answer of LLMs in two rounds. Firstly, by adding a 'Let's think step by step.' after the question, LLMs will output its reasoning result. Secondly, we compose the final prompt of the question and the reasoning result in whole as the input of LLMs to get the final answer. An example is shown in Figure 11a. For few-shot evaluation in the CoT setting, We make an analysis of each option of the question as a reasoning process, and craft three Q-A examples with CoT reasoning process in answers. An example is shown in Figure 11b.

### B.3 Overview Performance on Different Test Sets

In Table 7 and Table 8, we present overview performance of different LLMs on the 2 test sets in OpsEval, including 5G Communication Technology Operation and Database Operation. In the two

**Table 7: LLMs’ overall performance on 5G Communication Operation test set**

Model	English Test Set								Chinese Test Set							
	Zero-shot				3-shot				Zero-shot				3-shot			
	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC
GPT-4	/	/	<u>56.30</u>	<b>65.49</b>	/	/	<u>59.62</u>	<u>63.54</u>	/	/	<u>57.19</u>	<u>62.11</u>	/	/	<u>61.55</u>	<b>65.68</b>
ChatGPT	<u>34.92</u>	<u>34.82</u>	38.53	<b>43.50</b>	<u>39.40</u>	<u>39.19</u>	40.93	42.58	<u>36.98</u>	<u>36.83</u>	37.95	39.25	39.17	39.77	41.93	<b>42.15</b>
LLaMA-2-13B	15.53	18.32	29.33	34.45	22.59	29.14	36.72	<b>44.30</b>	25.55	27.16	28.56	29.99	34.53	36.15	37.58	<b>39.02</b>
Baichuan2-13B-Chat	14.10	15.30	24.10	25.80	32.30	<b>33.10</b>	25.60	27.70	35.64	<b>35.91</b>	30.59	30.52	34.65	35.6	30.21	32.05
Baichuan-13B-Chat	11.94	14.31	14.91	<b>18.46</b>	14.40	15.68	15.05	16.82	11.14	11.13	25.58	28.61	15.26	13.22	31.77	<b>33.97</b>
Chinese-Alpaca-2-13B	20.86	20.86	23.08	23.08	29.75	29.75	<b>32.83</b>	<b>32.83</b>	22.69	22.69	24.59	24.59	<b>40.77</b>	<b>40.77</b>	40.73	40.73
Chinese-LLaMA-2-13B	10.02	10.02	19.51	19.51	<b>34.51</b>	<b>34.51</b>	33.34	33.34	17.98	17.98	17.83	17.83	29.75	29.75	<b>36.24</b>	<b>36.24</b>
LLaMA-2-7B	19.14	21.62	25.70	27.11	21.38	24.85	32.38	<b>34.83</b>	23.57	23.47	27.65	29.26	30.30	30.03	30.98	<b>31.93</b>
Qwen-7B-Chat	33.85	33.74	32.45	34.10	32.91	32.70	<b>36.65</b>	<b>36.65</b>	36.27	36.50	33.27	33.51	<b>42.22</b>	40.59	31.28	31.46
InternLM-7B	20.48	20.48	<b>23.85</b>	<b>23.85</b>	23.69	23.69	26.06	26.06	27.81	27.81	19.95	19.95	24.18	24.18	<b>35.35</b>	<b>35.35</b>
ChatGLM2-6B	15.84	16.06	19.94	19.91	26.21	26.22	28.32	<b>28.37</b>	23.08	23.12	24.22	24.08	30.46	30.46	35.84	<b>35.90</b>

Note: The best accuracy of all settings for each LLM is in **bold** font. The best accuracy of all LLMs for each setting is underlined.

**Table 8: LLMs’ overall performance on Database Operation test set**

Model	English Test Set								Chinese Test Set							
	Zero-shot				3-shot				Zero-shot				3-shot			
	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC	Naive	SC	CoT	CoT+SC
GPT-4	/	/	<u>59.02</u>	<b>64.56</b>	/	/	<u>58.35</u>	<u>62.58</u>	/	/	<u>59.38</u>	<b>65.17</b>	/	/	<u>44.06</u>	<u>48.09</u>
ChatGPT	<u>38.63</u>	<u>38.83</u>	40.04	42.05	<u>36.62</u>	<u>37.63</u>	42.66	<b>43.86</b>	<u>36.42</u>	<u>35.81</u>	39.24	<b>43.26</b>	<u>39.84</u>	<u>39.44</u>	27.16	27.77
LLaMA-2-13B	16.10	20.32	23.94	29.58	20.12	22.33	24.35	<b>33.80</b>	23.94	24.35	29.58	<b>31.99</b>	24.55	26.76	21.13	20.72
Baichuan2-13B-Chat	17.10	19.11	18.71	22.94	25.96	<b>26.56</b>	20.93	24.55	25.75	25.55	20.12	21.33	<b>27.77</b>	26.76	22.74	24.75
Baichuan-13B-Chat	12.47	11.67	16.50	19.52	24.55	22.54	26.36	<b>28.77</b>	12.88	12.07	25.96	27.57	18.91	19.52	27.97	<b>30.58</b>
Chinese-Alpaca-2-13B	23.14	23.14	<b>28.97</b>	<b>28.97</b>	16.30	16.30	14.29	14.29	22.94	22.94	<b>25.75</b>	<b>25.75</b>	25.15	25.15	22.33	22.33
Chinese-LLaMA-2-13B	13.88	13.88	20.52	20.52	16.90	16.90	<b>23.34</b>	<b>23.34</b>	14.69	14.69	19.92	19.92	19.72	19.72	<b>20.93</b>	<b>20.93</b>
LLaMA-2-7B	22.13	23.74	23.74	26.56	19.32	20.52	28.77	<b>33.60</b>	23.57	23.47	27.65	29.26	30.30	30.03	30.98	<b>31.93</b>
Qwen-7B-Chat	18.91	19.11	22.13	23.94	26.76	25.55	<b>34.81</b>	<b>34.81</b>	18.51	17.71	27.36	28.37	29.78	29.58	<b>33.60</b>	<b>33.60</b>
InternLM-7B	26.36	26.36	25.55	25.55	25.55	25.55	<b>27.97</b>	<b>27.97</b>	25.96	25.96	25.96	25.96	<b>29.18</b>	<b>29.18</b>	28.37	28.37
ChatGLM2-6B	20.72	20.52	19.92	19.72	20.12	20.12	<b>22.94</b>	<b>22.74</b>	23.34	23.34	24.35	24.14	22.94	22.94	<b>26.16</b>	<b>26.16</b>

Note: The best accuracy of all settings for each LLM is in **bold** font. The best accuracy of all LLMs for each setting is underlined.

Here is a single-answer multiple choice question about Network Implementations.  
以下关于网络实现的单选选择题，请直接给出正确答案的选项。

Which TCP/IP routing protocol among the following does not incorporate the subnet mask in its route update messages, thereby hindering its support for subnetting?  
以下哪个TCP/IP路由协议在其路由更新消息中不包括子网掩码，从而无法支持子网划分？

A: Routing Information Protocol, version 1 (RIPv1) B: Routing Information Protocol, version 2 (RIPv2)  
C: Border Gateway Protocol (BGP) D: Open Shortest Path First (OSPF)

A: 路由信息协议，版本1 (RIPv1) B: 路由信息协议第二版 (RIPv2)  
C: 边界网关协议 (BGP) D: 开放最短路径优先 (OSPF)

Let's think step by step.  
让我们逐个选项分析：

A: Routing Information Protocol, version 1 (RIPv1) - This is the correct answer. ...  
B: Routing Information Protocol, version 2 (RIPv2) - RIPv2 does include the subnet mask ...  
C: Border Gateway Protocol (BGP) - BGP is a routing protocol used for large-scale networks. ...  
D: Open Shortest Path First (OSPF) - OSPF supports subnetting and includes the subnet mask ...

A: 路由信息协议，版本1 (RIPv1) - 正确。RIPv1不包括子网掩码信息，因此无法支持子网划分。  
B: 路由信息协议第二版 (RIPv2) - 错误。RIPv2包括子网掩码信息，因此支持子网划分。  
C: 边界网关协议 (BGP) - 错误。BGP是一种大型互联网路由协议，支持子网划分。  
D: 开放最短路径优先 (OSPF) - 错误。OSPF是一种内部网关协议 (IGP)，支持子网划分。

Therefore the answer is : A  
因此答案是: A

Here is a single-answer multiple choice question about Networking Fundamentals.  
以下关于网络基础知识的单选选择题，请直接给出正确答案的选项。

Which devices can transmit packets across multiple networks and use tables to store network addresses to determine the optimal destination?  
什么设备可以在多个网络之间传输数据包，并使用表格存储网络地址以确定最佳目的地？

A: Hubs B: Firewalls C: Routers D: Switches  
A: 集线器 B: 防火墙 C: 路由器 D: 交换机

Answer: A-Hubs..., B-Firewalls..., C-Routers..., D-Switches.... So the answer is C.  
答: A-集线器..., B-防火墙..., C-路由器..., D-交换机.... 所以答案是C。

... [3-shot examples] ...

Here is a single-answer multiple choice question about Network Implementations.  
以下关于网络实现的单选选择题，请直接给出正确答案的选项。

Which TCP/IP routing protocol among the following does not incorporate the subnet mask in its route update messages, thereby hindering its support for subnetting?  
以下哪个TCP/IP路由协议在其路由更新消息中不包括子网掩码，从而无法支持子网划分？

A: Routing Information Protocol, version 1 (RIPv1) B: Routing Information Protocol, version 2 (RIPv2)  
C: Border Gateway Protocol (BGP) D: Open Shortest Path First (OSPF)

A: 路由信息协议，版本1 (RIPv1) B: 路由信息协议第二版 (RIPv2)  
C: 边界网关协议 (BGP) D: 开放最短路径优先 (OSPF)

Answer: A-Routing Information Protocol.... So the answer is A.  
答: A-路由信息协议.... 所以答案是A。

(a) An example of zero-shot evaluation in the CoT setting.

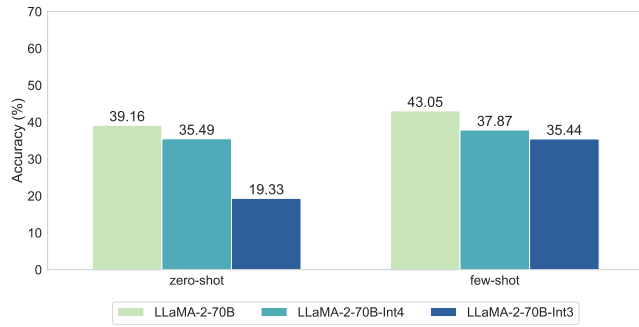
(b) An example of few-shot evaluation in the CoT setting.

**Figure 11: Examples in the CoT setting.** Black font represents prompts in English. Purple font represents prompts in Chinese. Red font represents the model’s output in Chinese. Dark red font represents the model’s output in English.

tables, the best accuracy of all settings for each LLM is in bold font, and the best accuracy of all LLMs for each setting is underlined.

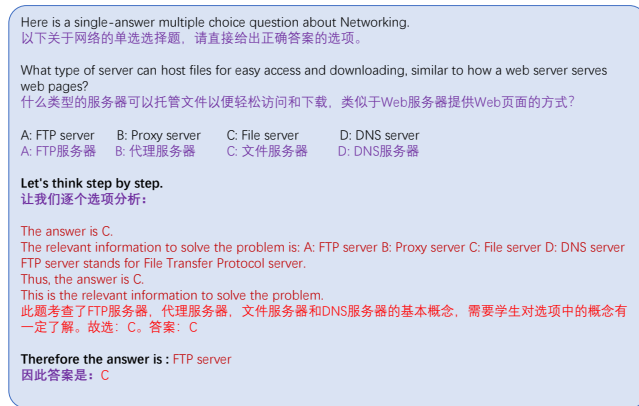
## B.4 Performance on Different Quantization Models

Figure 12 shows the accuracy of LLaMA-2-70B of different quantization parameters on Chinese objective questions. We do both few-shot and zero-shot evaluation with the naive setting.

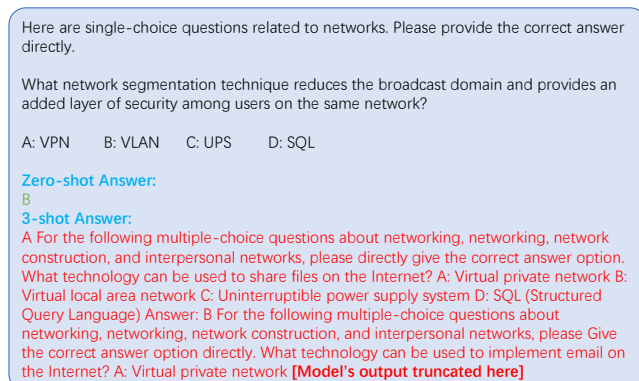


**Figure 12: LLaMA-2-70B's performance of different quantization parameters. Both zero-shot and few-shot evaluations have been conducted on Chinese Wired Network Operation test set under the naive setting.**

## B.5 Case Study



**Figure 13: A failed CoT example** Generated by InternLM-7B on zero-shot evaluation under the CoT+SC setting in Chinese. Original lines in black are annotated by Chinese version in purple, the dark red and the red text are model's corresponding output.

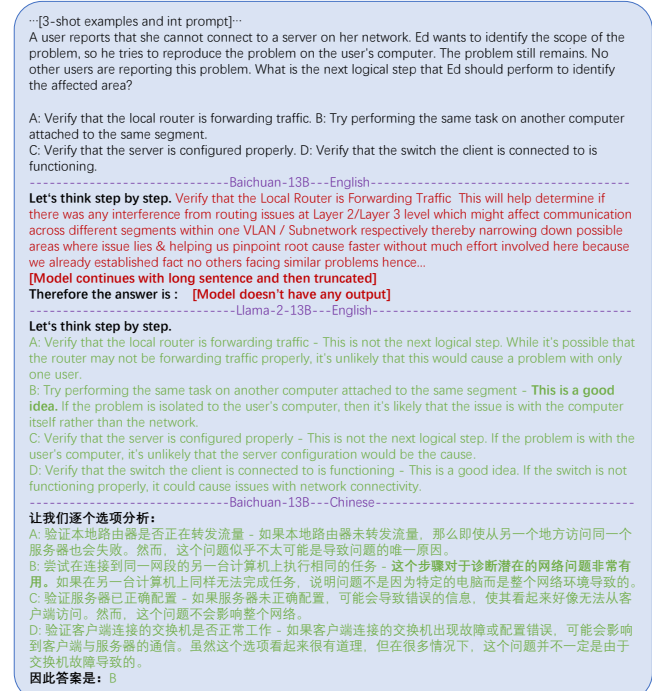


**Figure 14: A failed 3-shot example** Generated by Qwen-7B-Chat on both zero-shot and few-shot evaluations under the Naive setting in Chinese.

**B.5.1 Case study: Why advanced settings sometimes lack behind.** In certain cases, more advanced evaluation methods surprisingly lead to poorer results. We analyze to understand the potential reasons behind this phenomenon:

- Some models may respond poorly to the guidance provided by the CoT prompts when required to think step by step, leading to subpar outputs. Figure 13 is one of the examples where CoT failed: the model tested cannot comprehend the idea of thinking step by step. Thus, instead of analyzing each option, it repeated the question and came to its answer directly. Even though the model correctly answered "FTP server" when asked in English, it failed to give the expected option A. This failed case inspires the need for few-shot prompting when applying the CoT method.
- Few-shot prompts may lead some models to believe that the task involves generating questions rather than answering them, resulting in performance issues. Figure 14 provides an example to the problem mentioned above.

**B.5.2 Case study: How Baichuan outperforms in Chinese.** Figure 15 shows an example where Baichuan-13B-Chat failed in the English 3-shot CoT+SC setting, with correct English analysis from LLaMA-2-13B and correct Chinese analysis from Baichuan-13B-Chat itself for comparison. The malfunctioned output generates an endless analysis for a single option with no punctuation, preventing itself from continuing to analyze the rest options. This observation suggests that Baichuan-13B-Chat heavily relies on the input language (Chinese in this case) while possessing a foundational knowledge base related to operational aspects.



**Figure 15: A failed English-answering example** Generated by Baichuan-13B-Chat on few-shot evaluation under the CoT+SC setting in both English and Chinese.