Appendix

A Data sheet

We follow the documentation frameworks provided by Gebru et al. (2018).

A.1 Motivation

For what purpose was the dataset created?

The dataset was created to address the need for an effective benchmark to evaluate the performance of Large Language Models (LLMs) and Ops-specific LLMs (OpsLLMs) in IT operations (Ops) tasks. It aims to inform about the performance of current LLMs on Ops tasks and to aid in optimizing OpsLLMs tailored for the Ops domain. The benchmark, named OpsEval, was designed to tackle challenges such as sensitive data, numerous sub-domains, prompt sensitivity, and appropriate QA metrics in the Ops field.

A.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

The instances that comprise the dataset represent questions in two formats: multiple-choice questions (MC) and question-answering (QA) questions.

How many instances are there in total (of each type, if appropriate)?

7,184 multiple-choice questions and 1,736 QA questions.

What data does each instance consist of? Is there a label or target associated with each instance?

Multiple-choice questions consist of a stem, options, and an answer. Question-answering questions consist of a stem and an answer.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Questions in OpsEval are relatively independent, but they are clustered based on their classifications (8 tasks and 3 abilities) to group different questions by task type and general capabilities.

Are there recommended data splits (e.g., training, development/validation, testing)?

In our evaluation, we use few-shot evaluation. Therefore, we split the data into development (dev) and test sets. Detailed split information can be found in the dataset repository.

Is any information missing from individual instances? Are there any errors, sources of noise, or redundancies in the dataset? Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No information is missing from individual instances. No errors, sources of noise or redundancies in the dataset. The dataset is self-contained. The dataset does not contain offensive data.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?

The dataset does not contain data that might be considered confidential. We have desensitized the questions in the dataset to ensure that no internal private information from the participating companies is included.

A.3 Collection Process

How was the data associated with each instance acquired?

The data associated with each instance was acquired from four primary sources: company materials, certification exams, Ops textbooks, and automated generation.

Company materials include directly observable data such as Ops tickets and error logs, as well as internal documents and tests for Ops staff training, provided by cooperating companies from various sectors. Certification exam questions were sourced from public study guidebooks for Ops certification exams. Operations textbook data was acquired by searching for relevant books and extracting complete knowledge content and exercises. Automated generation involved using authoritative Ops textbooks and GPT4 to generate diverse questions. Each source is highly esteemed globally and reviewed by our Ops collaborators to ensure data validation and reliability.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

The dataset has been assembled from various sources through a combination of manual human curation and automated generation processes. More specifically:

- **Company Materials.** These materials were manually curated by experts from cooperating companies, including production environment materials like Ops tickets and error logs, as well as internal documents and tests for Ops staff training. Experts from 10 companies in sectors such as telecommunications, finance, and Ops service/tool providers contributed to this effort.
- **Certification Exams.** Certification exam questions were sourced from publicly available study guidebooks. These guidebooks, obtained from public book websites, contain knowledge assessments necessary for becoming an Ops staff and are naturally in the form of multiple-choice and question-answering questions. The questions were manually extracted from these guidebooks.
- **Operations Textbooks.** Relevant operations textbooks were identified by constructing a keyword list for the Ops field. These textbooks, which contain comprehensive knowledge content and exercises, were manually reviewed, and relevant questions were extracted for the dataset.
- **Automated Generation.** To enhance the diversity and depth of our test set, we used software programs to extract content from authoritative Ops textbooks and employed GPT-4 through software APIs to generate additional questions. This process involved manual verification by experts to ensure the quality and relevance of the generated questions.

The combination of these mechanisms ensures the reliability and robustness of the data for evaluating LLMs in the Ops domain. Full details, including the sources of the materials, are provided in the documentation accompanying our GitHub repository.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process involved a team comprising one undergraduate student, three graduate students, and over 20 experts from various companies. All participants, including the crowdworkers, voluntarily contributed to the data collection effort without any financial compensation.

Over what timeframe was the data collected?

The data was collected over the timeframe from July 2023 to May 2024, and the collection is still ongoing for the dataset's expansion and maintenance.

A.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Please refer to the OpsEval Benchmark section of the paper for details on any preprocessing, cleaning, or labeling of the data.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The "raw" data was saved in addition to the preprocessed, cleaned, and labeled data. However, it has not been made publicly available due to the inclusion of some internal company materials.

Is the software that was used to preprocess/clean/label the data available?

Yes, the scripts used for preprocessing, cleaning, and labeling the data are provided in the dataset repository.

A.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description.

The dataset has been used to evaluate the capabilities of large language models. For more details, please refer to the paper.

What (other) tasks could the dataset be used for?

The dataset could be used for various tasks, including evaluating the performance of Large Language Models (LLMs) and Ops-specific LLMs (OpsLLMs) in IT operations, such as network configuration, error log analysis, and operational knowledge assessments.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

Given the sensitivity and proprietary nature of some of the source materials, dataset consumers should be cautious about the potential legal and ethical implications of using this data. For example, some data might inadvertently reflect internal company processes or proprietary information.

To mitigate these risks, dataset consumers should:

- Use the data responsibly: Ensure that the data is used only for research and evaluation purposes and not for commercial exploitation.
- Avoid unfair treatment: Be mindful of potential biases in the data that could lead to stereotyping or unfair treatment of individuals or groups.
- Acknowledge data limitations: Recognize and disclose any limitations or biases in the data when publishing results or deploying models trained on this dataset.

Are there tasks for which the dataset should not be used?

1. Commercial purposes: Since some data is derived from proprietary company materials, commercial use could result in legal and ethical issues.
2. Sensitive decision-making processes: Avoid using the dataset for making decisions that could significantly impact individuals or groups, such as hiring decisions, without thoroughly evaluating the fairness and bias in the data.

A.6 Distribution

How will the dataset be distributed? When will the dataset be distributed? Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is currently distributed in Huggingface and Github.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No third parties have imposed IP-based or other restrictions on the data associated with the instances. No export controls or other regulatory restrictions apply to the dataset or to individual instances.

A.7 Maintenance

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The owner/curator/manager of the dataset can be contacted via email. (Our email address will be released once the paper be accepted.)

Is there an erratum? If so, please provide a link or other access point.

Currently, there is no erratum. If any errors are found in the future, they will be updated on GitHub. We welcome users to raise issues on GitHub to point out any errors.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

The dataset will be updated at least monthly by the authors. Updates will be announced via GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?

The dataset does not relate to personal data, so there are no applicable limits on the retention of data associated with individual instances.

Will older versions of the dataset continue to be supported/hosted/maintained?

Older versions of the dataset will be maintained in the GitHub history. Relevant updates will be communicated to users via GitHub notifications.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

We will communicate with contributors and review and process the contributed data following the same curation process outlined in the paper.

B Details of OpsEval Benchmark

B.1 Information on the companies and experts participating in OpsEval

Table 1: Information of companies collaborating in OpsEval

Organization	Domain	URL
Bank of Shanghai	Financial IT	https://www.bosc.cn/zh/
Bizseer	Ops service/tool 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/tool provider	https://www.zabbix.com/
Inspur	Ops service/tool provider	https://www.inspur.com/
Total	11	

Table 1 shows the companies participating in the creation of OpsEval benchmark suite. Their industries include the Internet, telecommunications, cloud computing, finance, and securities, and each company has dispatched at least two experts to participate in the OpsEval work.

B.2 Dependance Filtering Keyword List

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']

B.3 Prompt for GPT-4 Categorization

I need your help in analyzing a multi-choice question, determine the domain and the task type it belongs to.
Domains: When classifying the domain, be specific, dive deeper into domains such as: Database/Network Operations
Task Types: For the task type, consider categories like: Monitoring and Alerts, Performance Optimization
Summary your response as JSON format: {"domain": "specific_domain", "task": "specific_task_type"}

Figure 1: The prompt for GPT-4 initial categorization

Figure 1 shows the prompt for GPT-4 initial categorization.

B.4 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.

B.5 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.

Which of the following represents quantifying data moved from one host to another within a specific time frame?
A: Reliability B: Response time
C: Throughput D: Jitter
Answer: C
Analysis: Throughput is the measure of data transferred from one host to another in a given amount of time
Task: Performance Optimization
Ability: Knowledge Recall

Which command enables a router to signal clients that they should acquire additional configuration details from a DHCPv6 server?
A: ipv6 nd ra suppress B: ipv6 dhcp relay destination
C: ipv6 address autoconfig D: ipv6 nd other-config-flag
Answer: D
Analysis: The **ipv6** nd other-config-flag** command is used to enable a router to inform clients that they need to get additional configuration information from a DHCPv6 server
Task: Automation Scripts
Ability: Analytical Thinking

Question: You receive a call from a user experiencing difficulties connecting to a new VPN. What is the initial step you should take?
A: Find out what has changed. B: Reboot the workstation.
C. Document the solution. D: Identify the symptoms and potential causes.
Answer: D
Analysis: Since this is a new connection, you need to start by troubleshooting and identify the symptoms and potential causes
Task: Fault Analysis and Diagnostics
Ability: Practical Application

Figure 2: Three examples of the processed questions

Figure 2 illustrates examples in our question set, shedding light on our classification methodology.

B.6 Prompt and Formatting of Questions

Figure 3 illustrates examples of the questions after our preprocessing pipeline.

B.7 An Example of Subjective Questions

A saved subjective question in OpsEval is presented in Figure 4, which contains not only the raw question but also its type of task.

As shown in Figure 5, we combine the task and ability of each question with the question itself as the prompt for LLMs.

B.8 Scoring Rubrics of Fluency in FAE-Score

As show in Figure 6, we asked the judge model and experts about the aspects of grammatical correctness, coherence and consistency, calrity of expression, style and tone appropriateness and answer completion of the models' responses.

B.9 Automated QA generation

During the data collection process, we have experimented automating question-answer generation. We first sampled the QA pairs and manually assessed their accuracy and domain relevance. Later, we used typical manual evaluation examples for few-shot learning, enabling GPT to evaluate QA

Which of the following represents quantifying data moved from one host to another within a specific time frame?
A: Reliability B: Response time
C: Throughput D: Jitter
Answer: C
Analysis: Throughput is the measure of data transferred from one host to another in a given amount of time
Task: Performance Optimization
Ability: Knowledge Recall

Which command enables a router to signal clients that they should acquire additional configuration details from a DHCPv6 server?
A: ipv6 nd ra suppress B: ipv6 dhcp relay destination
C: ipv6 address autoconfig D: ipv6 nd other-config-flag
Answer: D
Analysis: The **ipv6** nd other-config-flag** command is used to enable a router to inform clients that they need to get additional configuration information from a DHCPv6 server
Task: Automation Scripts
Ability: Analytical Thinking

Question: You receive a call from a user experiencing difficulties connecting to a new VPN. What is the initial step you should take?
A: Find out what has changed. B: Reboot the workstation.
C. Document the solution. D: Identify the symptoms and potential causes.
Answer: D
Analysis: Since this is a new connection, you need to start by troubleshooting and identify the symptoms and potential causes
Task: Fault Analysis and Diagnostics
Ability: Practical Application

Figure 3: Three examples of the processed questions

Question: You have a router interface with an IP address of 192.168.192.10/29. What is the broadcast address that the hosts on this LAN will utilize?
问题：路由器上有一个接口，IP地址为192.168.192.10/29。主机在这个局域网上使用的广播地址是什么？

Keypoint: 192.168.192.15
答案要点：192.168.192.15

Detailed Answer: A /29 (255.255.255.248) has a block size of 8 in the fourth octet. This means the subnets are 0, 8, 16, 24, and so on. 10 is in the 8 subnet. The next subnet is 16, so 15 is the broadcast address.
答案解析：/29（255.255.255.248）在第四个八位组有8个块大小。这意味着子网是0，8，16，24等等。10在8的子网中。下一个子网是16，所以15是广播地址。

Task: Network Configuration
任务：网络配置

Ability: Analytical Thinking
能力：推理

Figure 4: An example of the saved subjective questions

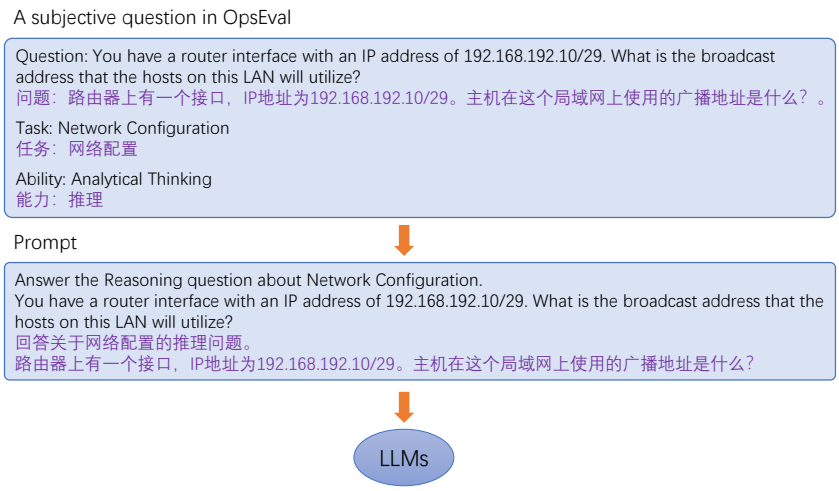

Figure 5: An example of building the prompt of subjective questions.

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 6: Scoring Rubrics of Fluency in FAE-Score.

Table 2: Models evaluated in this paper. The "access" column in the table shows whether we have full access to the model weights or can only access them through API.

Model	Creator	#Parameters	Access	License
GPT-4/GPT-3.5-turbo	OpenAI	*undisclosed*	API	Proprietary
ERNIE-Bot-4.0	Baidu	*undisclosed*	API	Proprietary
GLM4/GLM3-turbo	Tsinghua Zhipu	*undisclosed*	API	Proprietary
Meta-LLaMA-3	Meta	8B	Weights	Llama 3 Community
LLaMA-2	Meta	7/13/70B	Weights	Llama 2 Community
Qwen-Chat	Alibaba Cloud	7/14/72B	Weights	Qianwen LICENSE
Qwen1.5-Chat	Alibaba Cloud	14B	Weights	Qianwen LICENSE
InternLM2-Chat	Shanghai AI Laboratory	7/20B	Weights	Apache-2.0
DevOps-Model-Chat	CodeFuse	14B	Weights	Apache-2.0
Baichuan2-Chat	Baichuan Intelligence	13B	Weights	Apache-2.0
ChatGLM3	Tsinghua Zhipu	6B	Weights	Apache-2.0
Mistral	Mistral	7B	Weights	Apache-2.0
Gemma	Google	2/7B	Weights	Gemma license
Claude-3-Opus	Anthropic	*undisclosed*	API	Proprietary
Qwen2-Instruct	Alibaba Cloud	7/72B	Weights	Qianwen LICENSE

pairs based on our evaluation criteria automatically. Directly generated question-answers tend to be simple judgment or concept questions rather than reasoning questions that better demonstrate the model's capabilities and knowledge density. Our goal is to ensure that while the topics of the questions remain relevant to the seed questions, their specific content is distinct from the original questions. By maintaining the overarching framework in the Ops domain, we can expand the number and types of questions, enabling a more comprehensive evaluation of model capabilities. Additionally, we can incorporate external knowledge during the data generation, continually enhancing our ability to evaluate new content.

C Additional details of experiments

C.1 Detailed Information of LLMs Evaluated

GPT-4 (OpenAI, 2023) 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 lanuage model currently. ChatGPT (OpenAI, 2022) is an earlier AI-powered language model

Table 3: 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.

developed by OpenAI which is built upon GPT-3.5. We use the GPT-3.5-turbo version in our experiments. LLaMA 2 (Touvron, et.al., 2023) 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 2) of LLaMA 2.

Although LLaMA 2 is able to process Chinese input, it has a small Chinese vocabulary so that its abitilty of understanding and generating Chinese text is limited. As a result, we evaluate some Chinese-oriented LLMs which are published by institutions in China. ERNIE-Bot 4.0 (202, 2024) is the latest self-developed language model released by Baidu. As claimed by Baidu, ERNIE-Bot 4.0 rivals OpenAI's GPT-4. Qwen (202, 2023) (abbr. Tongyi Qianwen) is a series of LLMs developed by Alibaba Cloud. And Qwen-Chat is a series of large-model-based AI assistant trained with alignment techniques based on the pretrained Qwen. We evaluate three weights (72B, 14B and 7B as shown in 2) of Qwen-Chat. Baichuan2-13B-Chat (Baichuan, 2023) is aligned chat model based on Baichuan2-13B-Base (Baichuan, 2023) which is an open-source LLM published by Baichuan Intelligence. GLM (Du et al., 2022), developed by Tsinghua Knowledge Engineering Group, is a General Language Model pretrained with an autoregressive blank-filling objective and can be finetuned on various natural language understanding and generation tasks. Based on GLM, Zhipu AI released GLM4 (the newest version of GLM model) (Zeng et al., 2022) and GLM3 (the third version of GLM model). For GLM3, we use GLM3-turbo (Zeng et al., 2022) version and ChatGLM3-6B (Zeng et al., 2022) in our experiments. InternLM2-Chat-20B and InternLM2-Chat-7B (InternLM_Team, 2023), recently developed by Shanghai AI Laboratory, are multi-lingual models based on billions of parameters through multi-stage progressive training on over trillions of tokens. Furthermore, we evaluate DevOps-Model-14B-Chat (AI, 2024), an open source Chinese DevOps oriented models, mainly dedicated to exerting practical value in the field of DevOps.Gemma (Gemma_Team et al., 2024) is a family of lightweight, state-of-the-art open models based on Gemini technology from Google DeepMind. Trained on up to 6T tokens, Gemma achieves excellent language understanding and reasoning capabilities. We conducted an evaluation of Gemma-2b and Gemma-7b to investigate the effectiveness of Gemma with different weights.

In general, since some models (among them GPT-4, GPT-3.5-turbo, ERNIE-Bot-4.0, GLM4, GLM3-turbo) are not locally available, we evaluate them via API calls. For the remaining models, we perform local inference during evaluation.

C.2 Prompts

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 7. 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 8.

C.3 Compute and Resources Used for Experiments

During our OpEval experiments evaluating different LLMs, we utilize an 8 Nvidia A800-80GB GPU cluster to run inference on models with available weights. For models with API access, we perform inference using CPUs.

C.4 Overview Performance on Different Test Sets

In Table 4, Table 5 and Table 6, we present overview performance of different LLMs on the 3 test sets in OpsEval, including Wired Network Operations, 5G Communication Technology Operations and Database Operations.

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

Figure 7: An example of zero-shot evaluation 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.

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。

Figure 8: An example of few-shot evaluation 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.

Table 4: LLMs' overall performance on wired network operations 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	/	/	/	/	/	/	**88.70**	/	/	/	/	/	/	/	**86.00**	/
Qwen-72B-Chat	70.41	70.50	72.38	**72.56**	70.32	70.32	70.13	70.22	65.77	65.86	68.13	68.30	69.40	69.40	69.99	**70.08**
GPT-3.5-turbo	66.60	66.80	69.60	72.00	68.30	68.30	70.90	**72.50**	58.40	58.60	64.80	**67.60**	59.20	59.70	65.20	67.40
ERNIE-Bot-4.0	61.15	61.15	70.00	70.00	60.00	60.00	70.00	**70.00**	67.54	67.54	71.96	71.96	72.00	72.00	78.00	**78.00**
Qwen1.5-14B-Chat	54.90	34.88	64.09	60.82	52.23	**65.55**	59.54	47.08	54.04	45.18	62.56	59.12	58.78	61.10	**63.43**	52.5
Devops-Model-14B-Chat	30.69	30.59	55.77	63.63	**63.85**	61.96	41.15	44.01	47.59	46.57	52.52	56.01	**62.07**	60.08	50.59	55.79
Qwen-14B-Chat	43.78	47.81	56.58	59.40	**62.09**	59.70	49.06	55.88	48.35	48.81	55.35	57.40	**58.53**	56.12	52.12	54.99
LLaMA-2-13B	41.80	46.50	53.10	58.70	53.30	53.00	56.80	**61.00**	29.70	31.60	51.60	**57.00**	39.60	38.90	48.00	50.60
Gemma-7B	25.09	25.09	50.86	50.86	**59.12**	**59.12**	50.77	50.77	31.58	31.58	47.59	47.59	34.68	34.68	**48.88**	**48.88**
LLaMA-2-70B-Chat	25.29	25.29	57.97	58.06	52.97	52.97	**58.55**	**58.55**	38.55	38.55	**57.49**	**57.49**	49.09	49.09	48.57	48.57
Internlm2-Chat-20B	**56.36**	**56.36**	26.18	26.18	60.48	60.48	45.10	45.10	57.49	57.49	57.14	57.14	**59.12**	**59.12**	50.77	50.77
Internlm2-Chat-7B	49.74	49.74	**56.19**	**56.19**	48.20	48.20	49.74	49.74	57.49	57.49	57.14	57.14	**59.12**	**59.12**	50.77	50.77
LLaMA-2-7B	39.50	40.00	45.40	49.50	48.20	46.80	52.00	**55.20**	29.80	30.20	50.10	**55.60**	38.60	40.80	45.60	50.40
Qwen-7B-Chat	45.90	46.00	47.30	50.10	**52.10**	51.00	48.30	49.80	29.60	29.90	50.60	**53.50**	50.40	46.90	46.90	47.70
Baichuan2-13B-Chat	37.90	38.30	42.70	46.60	**51.90**	51.60	44.50	47.45	44.60	45.40	41.60	44.30	45.60	45.70	43.90	**46.70**

Note: The best accuracy of each language for each LLM is in **bold** font.

Table 5: LLMs' overall performance on 5G communication operations 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	/	/	56.30	**65.49**	/	/	59.62	63.54	/	/	57.19	62.11	/	/	61.55	**65.68**
Qwen-72B-Chat	53.19	53.19	55.25	55.52	58.13	58.13	58.72	**58.99**	64.79	64.79	**65.79**	65.72	70.19	70.19	68.31	68.38
InternLM2-Chat-20B	39.10	39.10	37.70	37.70	**47.70**	**47.70**	33.50	33.50	44.60	44.60	47.00	47.00	**62.20**	**62.20**	38.30	38.30
Qwen-14B-Chat	33.71	36.25	41.24	42.51	51.19	50.39	57.18	**59.18**	41.71	41.44	45.58	47.98	53.52	49.92	54.72	**58.85**
DevOps-Model-14B-Chat	31.04	30.51	42.84	47.37	**52.25**	49.38	45.90	47.23	41.04	42.70	48.71	53.57	56.85	**57.25**	51.30	54.29
ERNIE-Bot-4.0	43.66	43.66	**51.99**	**51.99**	44.00	44.00	50.00	50.00	45.99	45.99	48.98	48.98	46.00	46.00	**54.00**	**54.00**
LLaMA-2-70B	23.64	23.64	39.31	39.31	38.98	39.12	**47.90**	**47.90**	24.38	24.38	43.63	43.63	44.65	44.65	**48.84**	**48.84**
Mistral-7B	26.91	26.91	30.65	30.65	40.52	40.52	**46.84**	**46.84**	1.27	1.27	42.05	42.05	30.72	30.72	**46.44**	**46.44**
InternLM2-Chat-7B	36.80	36.80	31.70	31.70	**46.30**	**46.30**	36.90	36.90	38.80	38.80	44.60	44.60	**46.00**	**46.00**	35.80	35.80
Gemma-7B	23.10	23.10	**34.40**	**34.40**	21.40	21.40	33.10	33.10	27.30	27.30	35.40	35.40	17.30	17.30	**44.50**	**44.50**
LLaMA-2-13B	15.62	18.32	29.88	34.45	23.16	29.14	37.59	**44.3**	25.43	27.16	29.17	29.99	36.56	36.15	37.70	**39.02**
GPT-3.5-turbo	34.92	34.82	38.53	**43.50**	39.40	39.19	40.93	42.58	36.98	36.83	37.95	39.25	39.17	39.77	41.93	**42.15**
Qwen-7B-Chat	33.85	33.74	32.45	34.10	32.91	32.70	**36.65**	**36.65**	36.27	36.50	33.27	33.51	**42.22**	40.59	31.28	31.46
ChatGLM3-6B	30.40	30.40	30.70	30.70	26.90	26.90	**37.20**	**37.20**	32.60	32.60	35.40	35.40	28.30	28.30	**40.90**	**40.90**
Baichuan2-13B-Chat	14.10	15.30	24.10	25.80	32.30	**33.10**	25.60	27.70	35.64	**35.91**	30.59	30.52	34.65	35.6	30.21	32.05
LLaMA-2-7B	19.14	21.62	25.70	27.11	21.38	24.85	32.38	**34.83**	23.57	23.47	27.65	29.26	30.30	30.03	30.98	**31.93**
Gemma-2B	20.10	20.10	24.20	24.20	31.20	31.20	**35.50**	**35.50**	25.60	25.60	28.30	28.30	19.10	19.10	**35.50**	**35.50**

Note: The best accuracy of each language for each LLM is in **bold** font.

Table 6: LLMs' overall performance on database operations 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	/	/	59.02	**64.56**	/	/	58.35	62.58	/	/	59.38	**65.17**	/	/	44.06	48.09
InternLM2-Chat-20B	/	/	**59.21**	**59.21**	/	/	/	/	/	/	/	/	/	/	/	/
ERNIE-Bot-4.0	43.80	43.80	47.14	47.14	46.00	46.00	**54.0**	**54.0**	48.56	48.56	50.64	50.64	48.00	48.00	**54.0**	**54.0**
Gemma-7B	14.29	14.29	30.99	30.99	2.60	2.60	**43.86**	**43.86**	19.32	19.32	**53.95**	**53.95**	18.51	18.51	5.20	5.20
Qwen-72B-Chat	47.28	47.48	48.09	48.09	**49.70**	**49.70**	43.46	43.66	48.29	48.49	49.50	**49.70**	**49.70**	**49.70**	45.27	44.87
GPT-3.5-turbo	38.63	38.83	40.04	42.05	36.62	37.63	42.66	**43.86**	36.42	35.81	39.24	**43.26**	39.84	39.44	27.16	27.77
Qwen-14B-Chat	24.95	28.37	33.00	**36.62**	27.97	28.37	27.97	24.14	27.57	27.57	32.39	36.02	**40.04**	35.41	30.38	33.40
DevOps-Model-14B-Chat	25.15	26.96	35.41	**38.83**	33.20	34.81	27.36	27.36	24.75	22.74	28.37	27.77	36.62	**37.02**	27.57	26.36
LLaMA-2-70B	19.72	19.72	27.97	27.97	26.56	26.56	**32.6**	**32.6**	15.29	15.29	**34.81**	**34.81**	26.76	26.76	33.80	33.80
Qwen-7B-Chat	18.91	19.11	22.13	23.94	26.76	25.55	**34.81**	**34.81**	18.51	17.71	27.36	28.37	29.78	29.58	**33.60**	**33.60**
LLaMA-2-13B	16.10	20.32	23.94	29.58	20.12	22.33	24.35	**33.80**	23.94	24.35	29.58	**31.99**	24.55	26.76	21.13	20.72
LLaMA-2-7B	22.13	23.74	23.74	26.56	19.32	20.52	28.77	**33.60**	20.72	20.72	27.16	**27.97**	21.53	18.51	18.31	17.91
Mistral-7B	17.10	17.10	26.76	26.76	**31.19**	**31.19**	27.97	27.97	0.20	0.20	26.76	26.76	10.26	10.26	**32.19**	**32.19**
InternLM2-Chat-7B	27.16	27.16	28.17	28.17	29.98	29.98	**30.18**	**30.18**	28.57	28.57	**31.79**	**31.79**	30.78	30.78	31.19	31.19
ChatGLM3-6B	20.93	20.93	25.15	25.15	24.75	24.75	**29.18**	**29.18**	21.33	21.33	28.97	28.97	21.73	21.73	**29.58**	**29.58**
Baichuan2-13B-Chat	17.10	19.11	18.71	22.94	25.96	**26.56**	20.93	24.55	25.75	25.55	20.12	21.33	**27.77**	26.76	22.74	24.75
Gemma-2B	16.90	16.90	19.52	19.52	16.10	16.10	**24.75**	**24.75**	18.51	18.51	24.95	24.95	21.53	21.53	**27.77**	**27.77**

Note: The best accuracy of each language for each LLM is in **bold** font.

C.5 Performance on Different Quantization Models

Figure 9 shows the accuracy of LLaMA-2-70B of different quantization parameters on objective questions, English and Chinese questions respectively. We do both zero-shot and few-shot evaluation with the naive setting.

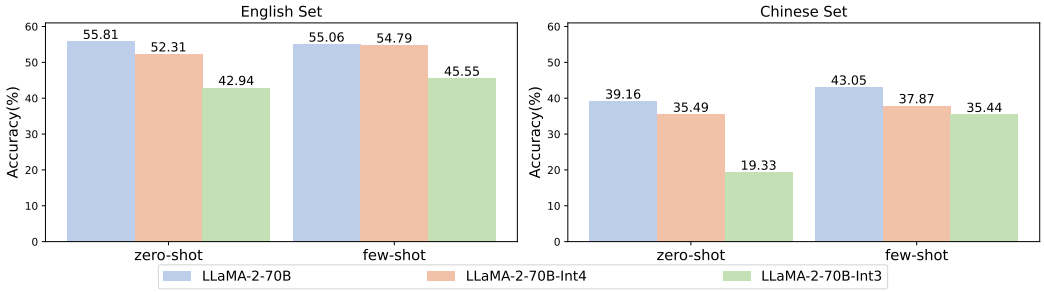

Figure 9: LLaMA-2-70B's performance of different quantization parameters. Both zero-shot and few-shot evaluations have been conducted on Wired Network Operations test set under the naive setting.

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

C.6 Performance on Different Languages

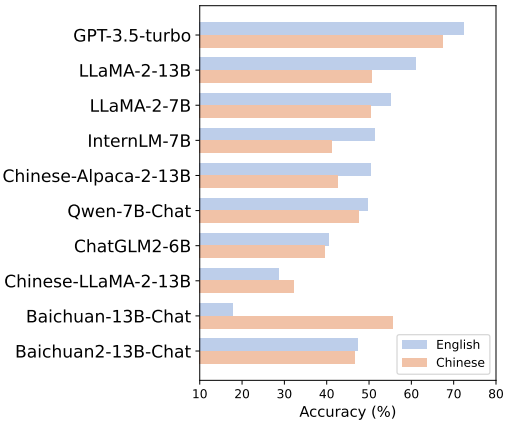

Figure 10: LLMs' few-shot performance on English/Chinese test set (CoT+SC)

In Figure 10, 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 as Chinese-Alpaca-2-13B, Qwen-7B-Chat, and ChatGLM2-6B, still perform better in answering English questions than Chinese ones.

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 C.9.4.

C.7 Expert alignment of FAE-Score

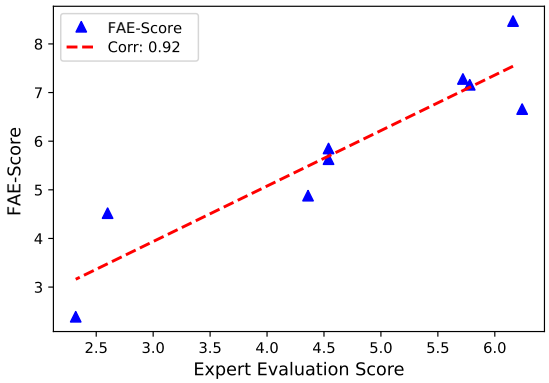

Figure 11: Scatter plot and trendline of FAE-Score compared to Expert Evaluation score.

As depicted in Figure 11, the FAE-Score demonstrates a strong positive correlation with Expert Evaluation Score, making it a valuable and effective substitute for automated evaluation.

C.8 Leakage Test Example

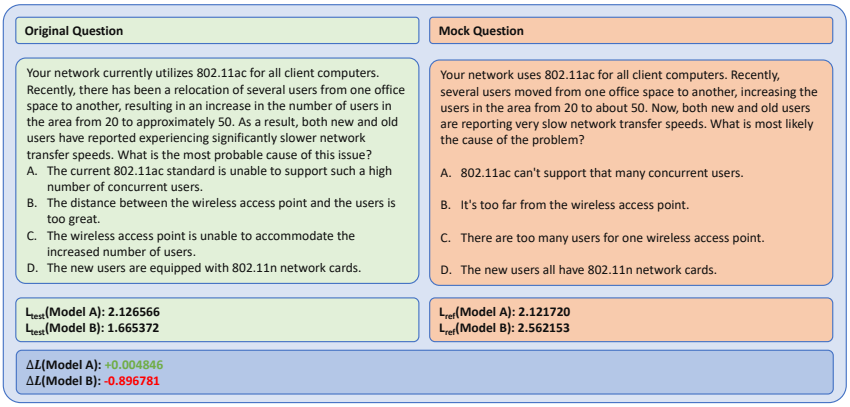

Figure 12: An example for leakage Test.

Figure 12 shows an example for leakage test. Note that the mock question is a paraphrase of the original question, but uses different vocabulary and phrasing. The difference in ΔL for Model B suggests that there may be potential leakage in the original question.

C.9 Case Study

C.9.1 Failure cases of OpsEval questions

Figure 13 illustrates three instances of failure in certain LLMs. The primary reasons for these failures are typically attributed to either a deficiency in domain knowledge or a lack of comprehension of domain knowledge.

C.9.2 Case analysis on QA metrics

Figure 14 illustrates an example where BLEU/ROUGE fail to pick up important keywords and are influenced by the repeated pattern attack, while FAE-Score avoids these two issues.

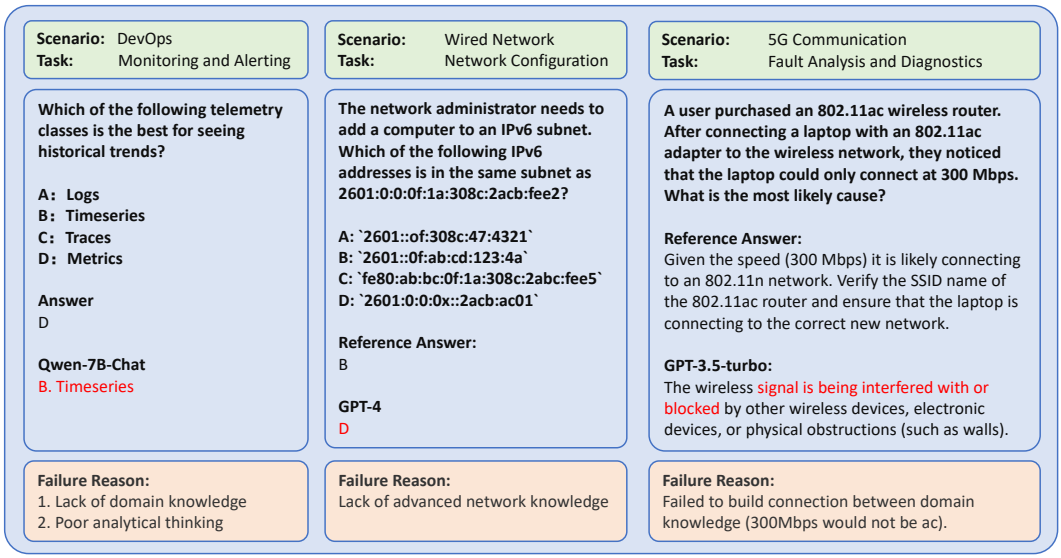

Figure 13: LLMs' failure cases of OpsEval questions.

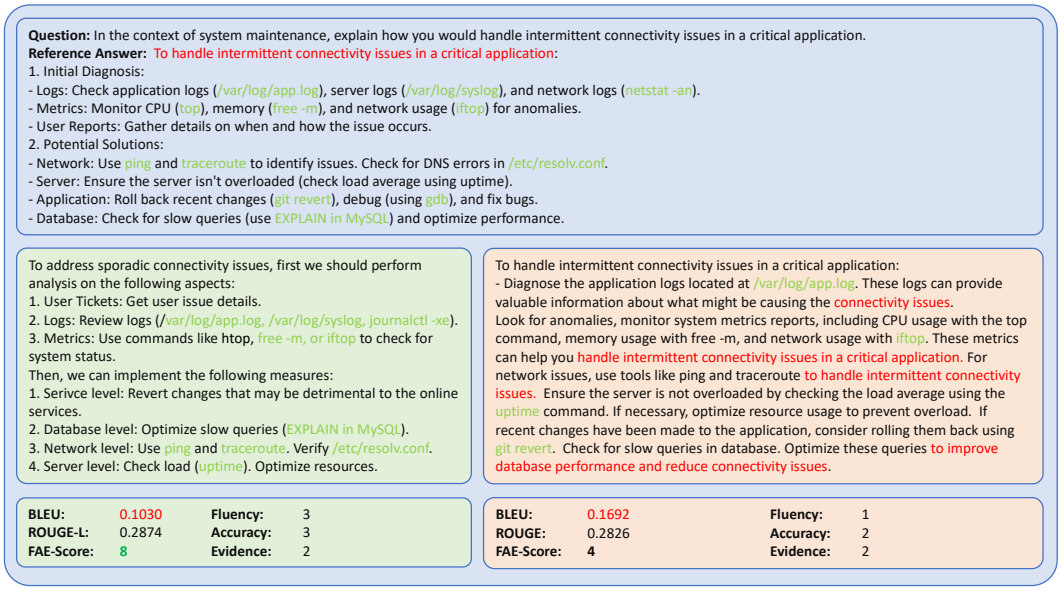

Figure 14: Case analysis on QA metrics

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

What type of server can host files for easy access and downloading, similar to how a web server serves web pages?
什么类型的服务器可以托管文件以便轻松访问和下载，类似于Web服务器提供Web页面的方式？

A: FTP server B: Proxy server C: File server D: DNS server
A: FTP服务器 B: 代理服务器 C: 文件服务器 D: DNS服务器

Let's think step by step.
让我们逐个选项分析：

The answer is C.
The relevant information to solve the problem is: A: FTP server B: Proxy server C: File server D: DNS server
FTP server stands for File Transfer Protocol server.
Thus, the answer is C.
This is the relevant information to solve the problem.
此题考查了FTP服务器，代理服务器，文件服务器和DNS服务器的基本概念，需要学生对选项中的概念有一定了解。故选：C。答案：C

Therefore the answer is : FTP server
因此答案是：C

Figure 15: 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.

Here are single-choice questions related to networks. Please provide the correct answer directly.

What network segmentation technique reduces the broadcast domain and provides an added layer of security among users on the same network?

A: VPN B: VLAN C: UPS D: SQL

Zero-shot Answer:
B
3-shot Answer:
A For the following multiple-choice questions about networking, networking, network construction, and interpersonal networks, please directly give the correct answer option. What technology can be used to share files on the Internet? A: Virtual private network B: Virtual local area network C: Uninterruptible power supply system D: SQL (Structured Query Language) Answer: B For the following multiple-choice questions about networking, networking, network construction, and interpersonal networks, please Give the correct answer option directly. What technology can be used to implement email on the Internet? A: Virtual private network **[Model's output truncated here]**

Figure 16: A failed 3-shot example Generated by Qwen-7B-Chat on both zero-shot and few-shot evaluations under the naive setting in Chinese.

C.9.3 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. 1) 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 15 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. 2) Few-shot prompts may lead some models to believe that the task involves generating questions rather than answering them, resulting in performance issues. Figure 16 provides an example to the problem mentioned above.

C.9.4 Case study: How Baichuan outperforms in Chinese

Figure 17 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 Ops.

···[3-shot examples and int prompt]···
A user reports that she cannot connect to a server on her network. Ed wants to identify the scope of the problem, so he tries to reproduce the problem on the user's computer. The problem still remains. No other users are reporting this problem. What is the next logical step that Ed should perform to identify the affected area?

A: Verify that the local router is forwarding traffic. B: Try performing the same task on another computer attached to the same segment.
C: Verify that the server is configured properly. D: Verify that the switch the client is connected to is functioning.

----------Baichuan-13B---English----------

Let's think step by step. Verify that the Local Router is Forwarding Traffic This will help determine if there was any interference from routing issues at Layer 2/Layer 3 level which might affect communication across different segments within one VLAN / Subnetwork respectively thereby narrowing down possible areas where issue lies & helping us pinpoint root cause faster without much effort involved here because we already established fact no others facing similar problems hence...
[Model continues with long sentence and then truncated]
Therefore the answer is : [Model doesn't have any output]

----------Llama-2-13B---English----------

Let's think step by step.
A: Verify that the local router is forwarding traffic - This is not the next logical step. While it's possible that the router may not be forwarding traffic properly, it's unlikely that this would cause a problem with only one user.
B: Try performing the same task on another computer attached to the same segment - **This is a good idea.** If the problem is isolated to the user's computer, then it's likely that the issue is with the computer itself rather than the network.
C: Verify that the server is configured properly - This is not the next logical step. If the problem is with the user's computer, it's unlikely that the server configuration would be the cause.
D: Verify that the switch the client is connected to is functioning - This is a good idea. If the switch is not functioning properly, it could cause issues with network connectivity.

----------Baichuan-13B---Chinese----------

让我们逐个选项分析：
A: 验证本地路由器是否正在转发流量 - 如果本地路由器未转发流量，那么即使从另一个地方访问同一个服务器也会失败。然而，这个问题似乎不太可能是导致问题的唯一原因。
B: 尝试在连接到同一网段的另一台计算机上执行相同的任务 - **这个步骤对于诊断潜在的网络问题非常有用。**如果在另一台计算机上同样无法完成任务，说明问题不是因为特定的电脑而是整个网络环境导致的。
C: 验证服务器已正确配置 - 如果服务器未正确配置，可能会导致错误的信息，使其看起来好像无法从客户端访问。然而，这个问题不会影响整个网络。
D: 验证客户端连接的交换机是否正常工作 - 如果客户端连接的交换机出现故障或配置错误，可能会影响到客户端与服务器的通信。虽然这个选项看起来很有道理，但在很多情况下，这个问题并不一定是由于交换机故障导致的。
因此答案是： B

Figure 17: A failed English-answering example Generated by Baichuan-13B-Chat on few-shot evaluation under the CoT+SC setting in both English and Chinese.

D Annotation Guideline for OpsEval Categorization

D.1 Overview

In the OpsEval project, we aim to categorize operational and maintenance tasks within the industry. This categorization process is pivotal for understanding the spectrum of tasks and the required abilities to address them effectively. The process involves two primary steps: automated screening using GPT-4 for initial topic modeling, followed by a manual review process involving domain experts.

D.2 Task Categorization

D.2.1 Objective

To categorize questions into one of eight distinct operational tasks based on industry relevance, task frequency, and significance within operational settings.

D.2.2 Steps

1. **Review Initial Categorization**: Begin with the insights provided by GPT-4's topic modeling. Each question has been preliminarily categorized into one or more operational tasks.

2. **Understand Task Definitions**: Familiarize yourself with the details of the eight distinct tasks outlined in the provided Appendix. Each task has specific criteria and examples to guide your categorization.
3. **Assign Tasks**: For each question, decide which of the eight tasks it belongs to. A question should be categorized based on its core focus and the operational activity it pertains to.
4. **Justification**: Briefly justify your choice, especially if a question seems to fit into more than one category. Use the task definitions as a guide to support your decision.

D.2.3 Detailed Task Categorizations

1. **General Knowledge**: Questions related to foundational concepts and practices in the Ops domain.
2. **Fault Analysis and Diagnostics**: Questions focusing on identifying and solving discrepancies or faults in systems or networks.
3. **Network Configuration**: Questions about optimal configurations for network devices to ensure efficient and secure operations.
4. **Software Deployment**: Questions dealing with the distribution and management of software applications.
5. **Monitoring and Alerts**: Questions on using monitoring tools to oversee system efficiency and setting up alert mechanisms.
6. **Performance Optimization**: Questions aimed at enhancing network and system performance.
7. **Automation Scripts**: Questions involving the creation of scripts to automate processes and reduce manual intervention.
8. **Miscellaneous**: Questions that do not fit into the above categories or involve elements from multiple categories.

D.2.4 Task Categorization Template

Question ID:
Question: [Insert question text here]
Assigned Task:
Justification: [Provide a brief explanation for the task assignment here]

D.2.5 Example for Task Categorization

Question ID: 001
Question: What steps should be taken to configure a firewall to prevent unauthorized access while allowing legitimate traffic?
Assigned Task: Network Configuration
Justification: This question specifically asks for optimal configuration strategies for a key network device (firewall) to ensure security and efficient operation, aligning perfectly with the 'Network Configuration' task.

D.3 Ability Categorization

D.3.1 Objective

To classify questions based on the required cognitive ability to answer them: Knowledge Recall, Analytical Thinking, or Practical Application.

D.3.2 Steps

1. **Review Definitions**: Read the descriptions of the three abilities in the provided Appendix. Each ability category has distinct characteristics and examples.
2. **Evaluate Questions**: Assess the cognitive demand of each question. Consider what is primarily required to answer the question effectively: recalling information, analyzing data/situations, or applying knowledge in practical scenarios.

3. **Assign Ability Level**: Determine the most appropriate ability category for each question. Some questions may seem to require multiple abilities; choose the one that is most critical for addressing the core challenge of the question.
4. **Justification**: Provide a rationale for your categorization, especially for questions that may not clearly fit into a single category. Refer to the ability definitions to support your categorization.

D.3.3 Detailed Ability Categorizations

1. **Knowledge Recall**: Requires recognizing and recalling core concepts and foundational knowledge.
2. **Analytical Thinking**: Demands deeper thought to dissect problems, correlate information, and derive conclusions.
3. **Practical Application**: Involves applying knowledge or analytical insights to provide actionable recommendations.

D.3.4 Ability Categorization Template

Question ID:
Question: [Insert question text here]
Assigned Ability:
Justification: [Provide a brief explanation for the ability level assignment here]

D.3.5 Example for Ability Categorization

Question ID: 002 Question: How would you optimize the performance of a network experiencing frequent bottlenecks?
Assigned Ability: Practical Application Justification: The question requires applying knowledge of network systems and performance optimization techniques to propose specific solutions, hence it falls under 'Practical Application'.

D.4 General Guidelines

- **Consistency**: Strive for consistency in your categorization decisions. If similar questions are categorized differently, reassess your choices to ensure they align with the task and ability definitions.
- **Collaboration**: When in doubt, discuss challenging questions with fellow experts. Collaboration can help clarify ambiguities and refine the categorization process.
- **Documentation**: Keep detailed notes on your decisions, especially for questions that required significant deliberation. This documentation will be valuable for future reference and analysis.

By following these guidelines, you will contribute to a comprehensive and nuanced categorization of operational tasks and required abilities. This effort is crucial for enhancing our understanding of the operational landscape and the diverse skills professionals need to navigate it effectively.

References

QwenLM/Qwen-7B. IX 2023.

baidu/ERNIE-Bot-4.0. IX 2024.

AI CodeFuse. CodeFuse-DevOps-Model. 2024. Accessed: 2024-06-06.

Baichuan . Baichuan 2: Open Large-scale Language Models // arXiv preprint arXiv:2309.10305. 2023.

Du Zhengxiao, Qian Yujie, Liu Xiao, Ding Ming, Qiu Jiezhong, Yang Zhilin, Tang Jie. GLM: General Language Model Pretraining with Autoregressive Blank Infilling // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. 320–335.

Gebru Timnit, Morgenstern Jamie, Vecchione Briana, Vaughan Jennifer Wortman, Wallach Hanna M., III Hal Daumé, Crawford Kate. Datasheets for Datasets // CoRR. 2018. abs/1803.09010.

Gemma_Team , Mesnard Thomas, et.al. . Gemma: Open Models Based on Gemini Research and Technology. 2024.

InternLM_Team . InternLM: A Multilingual Language Model with Progressively Enhanced Capabilities. 2023.

OpenAI . ChatGPT: Optimizing Language Models for Dialogue // OpenAI Blog. 2022.

OpenAI . GPT-4 Technical Report // arXiv preprint arXiv:2303.08774. 2023.

Touvron Hugo, et.al. . Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023.

Zeng Aohan, Liu Xiao, Du Zhengxiao, Wang Zihan, Lai Hanyu, Ding Ming, Yang Zhuoyi, Xu Yifan, Zheng Wendi, Xia Xiao, others . Glm-130b: An open bilingual pre-trained model // arXiv preprint arXiv:2210.02414. 2022.