BIBench: Benchmarking Data Analysis Knowledge of Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of tasks. However, their proficiency and reliability in the specialized domain of Data Analysis, particularly with a focus on data-driven thinking, remain uncertain. To bridge this gap, we introduce BIBench, a comprehensive benchmark designed to evaluate the data analysis capabilities of LLMs within the context of Business Intelligence (BI). BIBench assesses LLMs across three dimensions: 1) BI foundational knowledge, evaluating the models' numerical reasoning and familiarity with financial concepts; 2) BI knowledge application, determining the models' ability to quickly comprehend textual information and generate analysis questions from multiple views; and 3) BI technical skills, examining the models' use of technical knowledge to address real-world data analysis challenges. BIBench comprises 11 sub-tasks, spanning three categories of task types: classification, extraction, and generation. Additionally, we've developed BIChat, a domain-specific dataset with over a million data points, to fine-tune LLMs. We will release BIBenchmark, BIChat, and the evaluation scripts at `https://github.com/xxx`. This benchmark aims to provide a measure for in-depth analysis of LLM abilities and foster the advancement of LLMs in the field of data analysis.

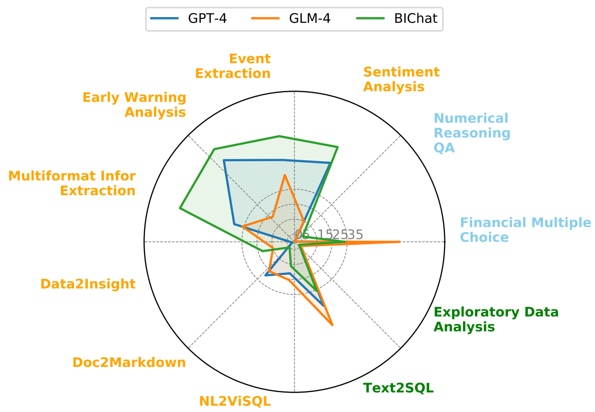

Figure 1: Results (zero-shot) of three best-performing LLMs evaluated on 11 diverse BI tests covering three cognitive dimensions.

1 Introduction

With the advance in pre-trained language models(Devlin et al., 2019), the Natural Language Processing (NLP) technology is evolving fast, so as its applications in finanical and business domains (Yang et al., 2020). With the release of ChatGPT series (OpenAI, 2022), decoder-only Large Language Models (LLMs) like GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b) have rapidly become a cornerstone of modern artificial intelligence, demonstrating remarkable versatility and power in NLP. Their ability to understand, generate, and sometimes even reason with human language has led to transformative applications across numerous fields (Huang et al., 2023; Zhong et al., 2023). However, despite their broad capabilities, the performance of LLMs in specialized domains, particularly those requiring data-driven analytical skills, has not been thoroughly examined.

Business Intelligence (BI) is one such domain where decision-making is deeply rooted in data analysis. It demands not only an understanding of complex financial and operational concepts but also the ability to apply this knowledge practically. This is much different from previous financial benchmarks like BBT-Fin (Lu et al., 2023), FinEval (Zhang et al., 2023) and PIXIU (Xie et al., 2023) where only financial concepts are evaluated via question-answering. BI involves synthesizing information from diverse sources, asking pertinent questions, and employing technical skills to navigate and interpret data. The nuanced requirements of BI pose a unique challenge for LLMs, which have typically been evaluated on more general language tasks.

To address this challenge, we introduce BIBench, a pioneering benchmark specifically designed to probe the depths of LLMs' data analysis capabilities within the Business Intelligence domain. BIBench aims to provide a multi-faceted evaluation framework that mirrors the multifarious nature of BI tasks. It is structured to assess LLMs in three critical dimensions: 1) BI Foundational Knowledge: This dimension tests the LLMs' grasp of numerical reasoning and their understanding of fundamental financial concepts. It is essential for models to exhibit a strong base knowledge of BI to ensure their reasoning is grounded in the realities of the domain. 2) BI Knowledge Application: Beyond foundational knowledge, LLMs must demonstrate the ability to parse through textual information swiftly and formulate analysis questions from multiple perspectives. This dimension simulates the real-world scenario where analysts derive meaningful questions that drive data exploration. 3) BI Technical Skills: The third dimension pushes LLMs to showcase their technical prowess in addressing real-world data analysis challenges. This includes the use of BI tools, understanding of data structures, and the ability to generate insights through analytical reasoning.

BIBench is composed of 11 sub-tasks, which fall under three categories of task types—classification, extraction, and generation. Together, these tasks constitute a comprehensive suite that rigorously tests the models across the spectrum of skills required in BI. In conjunction with BIBench, we have developed BIChat, a domain-specific dataset encompassing over a million data points to fine-tune LLMs. BIChat is finetuned from Qwen (Bai et al., 2023), a strong LLM base which has brilliant abilities in both English and Chinese, to ensure that the models are attuned to the nuances of language and concepts prevalent in the BI research. Our goal is to establish a standard for in-depth evaluation of LLMs in the context of BI and catalyze further advancement in the application of LLMs to data analysis. By doing so, we hope to bridge the gap between the capabilities of general-purpose LLMs and the specialized demands of BI, paving the way for more sophisticated and reliable AI tools in the realm of business and beyond.

The evaluation results are shown in Figure 1. We find that although BI specific fine-tuning usually improves upon their base model, they are still lagging behind general LLMs in some sub-tasks, which occupy the top three spots in the averaged zero-shot performance. We analyze the impact of various factors on the results, such as supervised fine-tuning (SFT), BI specific fine-tuning, etc.

Our contributions are summarized as follows:

- We construct BIBench, which includes 11 subtasks across three dimensions to evaluate the data analysis capabilities of LLMs.
- We systematically benchmarked 41 popular LLMs' data analysis capabilities for the first time. On top of their performance on BIBench, we provide deep insights into status quo of BI LLMs' development and highlight the deficiencies that require improvements
- We trained BIChat, which to our knowledge, is the first LLM tailored for the business intelligence domain.

2 Related Work

Benchmarks for Large Language Models The assessment of the capabilities and delimitation of the potential for LLMs is an area of avid scholarly pursuit. Previous studies have focused on assessing the overall proficiency or targeted-domain competencies in resolving practical challenges.

Regarding comprehensive capabilities, **HaluEval** (Li et al., 2023b) incorporates a dataset of 35,000 instances for the identification of errors in general conversations, Question Answering, Dialogue, and Summarization. The dataset known as **Felm** (Chen et al., 2023a) comprises 817 samples and 3,948 sections encompassing domains such as general knowledge, science and technology, literary and evaluative faculties, mathematics, and logic to scrutinize the factual accuracy of LLMs. Furthermore, **GSM8K** (Cobbe et al., 2021) presents a collection of 8,500 superior quality grade school mathematical challenges, crafted by expert problem creators, that require between two and eight steps to resolve, typically necessitating the execution of a series of elementary arithmetic operations (addition, subtraction, division, multiplication) to derive the conclusive resolution. The benchmark **MMLU** (Hendrycks et al., 2020) is engineered to ascertain the knowledge accrued during pre-training by examining models exclusively in the context of zero-shot and few-shot environments, spanning 57 subject areas across Science, Technology, Engineering, Mathematics (STEM), the humanities, social sciences, among others.

Cognitive Level	ID	Task	Data Source	Metric	Type
Foundational Knowledge	1-1	Financial Multiple Choice	FinExam	F1	Classification
	1-2	Numerical Reasoning QA	ConvFinQA	Accuracy	Generation
Knowledge Applications	2-1	Sentiment Analysis	CCKS2020	F1	Extraction
	2-2	Event Extraction	CCKS2020	F1	Extraction
	2-3	Early Warning Analysis	News	F1	Extraction
	2-4	Multiformat Infor Extraction	LIC2021	F1	Extraction
	2-5	Data2Insight	Spider	Rouge-L	Generation
	2-6	Doc2Markdown	FinGLM	Rouge-L	Generation
	2-7	NL2ViSQL	Spider	Rouge-L	Generation
Technical Skills	3-1	Text2SQL	BIRD	Rouge-L	Generation
	3-2	Exploratory Data Analysis	Kaggle.com	Rouge-L	Generation

Table 1: Task list in BIBench. Tasks correspond to cognitive dimensions: Foundational knowledge, knowledge Applcaitons and Technical Skills.

Additionally, concerning domain-specific competencies, **LawBench** (Fei et al., 2023) incorporates 20 tasks arranged across three cognitive proficiency tiers. **MultiMedQA** (Singhal et al., 2023) is composed of six healthcare-related Question Answering datasets devised for the assessment of clinical knowledge. In proximity to our research, **FinGPT** (Liu et al., 2023a) catalogs a limited selection of tasks such as Named Entity Recognition (NER) within the financial sector. Notwithstanding, our BIBench extends into a more granular analysis and is categorized into three cognitive stratifications.

Advancements in Large Language Models In the realm of computational linguistics, there has been a profound and accelerating interest in Large Language Models (LLMs), which are trained on vast textual corpora. These models have shown remarkable ability in generating high-quality text across a spectrum of applications, both general and domain-specific (Zhao et al., 2023; Min et al., 2023; Yang et al., 2023). LLMs can be classified into two categories based on their availability: closed-source and open-source models. Prominent examples of closed-source LLMs include the GPT-family[1], Claude[2], Gemini[3], and ERNIE[4]. Nevertheless, there has been an increasing focus on open-source LLMs that provide comprehensive access to their model weights, thereby facilitating deeper research exploration. A notable open-source LLM is LLaMA-2 (Touvron et al., 2023b), pioneered by Meta, which offers support for 20 languages and is an evolution of its precursor, LLaMA-1 (Touvron et al., 2023a). Additionally, the ChatGLM-family (Du et al., 2022; Zeng et al., 2022) represents multilingual models with prowess in both English and Chinese, among others. In this paper, we introduce an innovative benchmark designed to facilitate a more comprehensive investigation from the persepective of business intelligence of large language models.

[1]https://chat.openai.com
[2]https://claude.ai
[3]https://www.gemini.com
[4]https://yiyan.baidu.com

3 BI Benchmark

In this section, we describe in detail the design rationale of BIBench and the selected testing tasks.

3.1 The Taxonomy of BIBench

We have adopted Bloom's Taxonomy (Krathwohl, 2002), which provides a widely accepted framework for classifying tasks into different dimensions (Yu et al., 2023). Inspired by this classification approach, we have simplified Bloom's cognitive hierarchy model. In our approach, we focus on the first three categories of Bloom's Taxonomy to evaluate the BI knowledge of LLMs.

Foundational Knowledge: The fundamental knowledge level gauges the essential requisites of memorizing financial knowledge and numerical computation. It assesses LLMs' capacity for retaining elementary knowledge in the financial domain and their proficiency in executing multiple rounds of numerical calculations, encompassing concepts, general knowledge, financial facts, terminology,

Type	Model	Parameters	Instruction	RL	Access	BaseModel
English LLMs	GPT-4-0613	—	✓	✓	API	—
	GPT-3.5-turbo-0613	—	✓	✓	API	—
	LLaMA2-Base	7/13/70B	✓	✗	Weights	—
	LLaMA2-Chat	7/13/70B	✓	✓	Weights	LLaMA2-7/13/70B
	Vicuna-v1.5	7B	✓	✗	Weights	LLaMA2-7B
	Alpaca-v1.0	7B	✓	✗	Weights	LLaMA-7B
	WizardLM	7B	✓	✗	Weights	LLaMA-7B
	Phi	2B	✓	✗	Weights	—
Chinese LLMs	通义千问(Qwen-turbo)	—	✓	✓	API	—
	文心一言(ERNIEv4.0)	—	✓	✓	API	—
	智谱清言(GLM-4)	—	✓	✓	API	—
	Yi-Base	6B/34B	✓	✗	Weights	—
	Yi-Chat	6B/34B	✓	✗	Weights	Yi-6B/34B
	InternLM-Base	7B/20B	✓	✗	Weights	—
	InternLM-Chat	7B/20B	✓	✗	Weights	InternLM-7B
	Qwen-Base	7B/14B	✓	✗	Weights	—
	Qwen-Chat	1.8B/7B/14B	✓	✗	Weights	Qwen-1.8/7/14B
	Baichuan2-Base	7B/13B	✓	✗	Weights	—
	Baichuan2-Chat	7B/13B	✓	✗	Weights	Baichuan2-7/13B
	TigerBot-Base	7B	✓	✗	Weights	—
	TigerBot-Chat	7B	✓	✗	Weights	TigerBot-7B
	Chinese-Alpace2	7B	✓	✗	Weights	LLaMA2-7B
	ChatGLM2	6B	✓	✗	Weights	ChatGLM-6B
	ChatGLM3-Base	6B	✓	✗	Weights	—
	ChatGLM3	6B	✓	✗	Weights	ChatGLM3-6B-Base
	MiniCPM	2B	✓	✗	Weights	—
BI LLMs	XuanYuan-Chat	13/70B	✓	✗	Weights	LLaMA2-13/70B
	BIChat (ours)	7B	✓	✗	Weights	Qwen-7B-Chat

Table 2: LLMs tested on BIBench. We classify these models by their main training corpora.

and basic computational skills.

Knowledge Applications: The knowledge applications level includes understanding the meaning of financial documents and the perspective of data analysis. This includes the ability to understand concepts, texts and issues in finance and business, for example, to identify entities and events in texts, detect financial risks for data analysis, etc.

Technical Skills: The technical skills requires large language models to understand financial knowledge and problems, conduct data analysis and mining on them, and give in-depth suggestions for exploratory data analysis. It covers the ability of the model to understand text problems and give the SQL to solve them, as well as the depth of understanding to perform exploratory data analysis.

3.2 Selected Tasks and Datasets

We have chosen 11 tasks that correspond to the aforementioned capability levels, and each task has been assigned a unique identifier for better differentiation. The list of tasks can be found in Table 1. For every task, we have created a dataset of formations. When constructing BIBench, we have taken great care to format the prompts in a manner that is most consistent with user habits while providing clear instructions on answer formatting so as to evaluate LLMs' ability to assist with BI tasks under realistic conditions.

BI Foundational Knowledge Tasks

BI Foundational knowledge tasks aim to assess the extent to which large language models encode industry-specific knowledge and numerical computation within their parameters. There are two principal categories of knowledge that require mastery: (1) Fundamental financial knowledge, and (2) data computation. To evaluate these types of knowledge, we have devised two distinct tasks:

Financial Multiple Choice (1-1): *Given a question asking about basic financial knowledge, select the correct answer from 4 candidates.* We collected finance-related questions from the dataset of XuanYuan (Zhang and Yang, 2023), sampling 500 examples. These examples comprise both single-choice and multiple-choice questions.

Numerical reasoning QA (1-2): *Multiple rounds of numerical q&a calculations based on table content.* We use the ConvFinQA (Chen et al., 2022) dataset to build this task. We sampled 500 examples from the dataset.

Examples of these tasks are in Appendix A.1.

BI Knowledge Applications Tasks

BI knowledge Applications tasks examine to which extent large language models can comprehend entities, events and BI requirements in business scenarios. Understanding business scenarios is a precondition to utilize the knowledge in concrete downstream applications. In total, we selected seven tasks corresponding to different levels of BI knowledge Applications.

Sentiment Analysis (2-1): *Extract the subject and type of public opinion from the given news.* We sampled 600 examples from the CCKS2020_FEE_task1 [5] dataset as our test set. These 600 samples contain 11 entity types amrelated to public opinion.

Event Extraction (2-2): *Extract the event type and various event elements based on the text.* We randomly select 250 samples from the CCKS2020 Cross-category Event Extraction with Few-shot dataset [6] for this task. This dataset consists of 10 categories and their corresponding event elements.

Early Warning Analysis (2-3): *Extract the financial entities, along with their associated opportunity labels and risk labels, based on the content.* We have gathered 300 financial news articles and enlisted the expertise of relevant specialists to annotate each article with opportunity labels and risk labels. The labeling system includes 102 risk labels and 93 opportunity labels.

Multiformat Infor Extraction (2-4): *Given a sentence and predefined event types with corresponding argument roles, the aim of this task is to identify all events of target types mentioned in the sentence, and extract corresponding event arguments playing target roles.* We manually selected 250 samples from the LIC2021 DuEE-fin dataset[7] to serve as the test dataset, encompassing 13 event types. These documents originate from news and announcements in the financial domain, covering a wide range of problems encountered in real-world scenarios.

Data2Insight (2-5): *Generate data analysis suggestions and insights from the given structured data.* We constructed prompt instructions based on the query results of the Spider (Yu et al., 2019) dataset and employed a self-supervised approach using GPT-4 to generate data analysis insights. Subsequently, human verification was conducted, resulting in a test dataset of 300 samples.

Doc2Markdown (2-6): *Convert essential information from financial texts into a Markdown table.* We employed the FinGLM financial report [8] and utilized the PDF parsing tool pdfumber to extract relevant financial information. We created prompt instructions to generate corresponding Markdown tables using GPT-4. After manual verification, we retained 300 samples as the test dataset.

NL2ViSQL (2-7): *Generate SQL analysis statements from given questions and table structures, considering multiple perspectives.* We utilized the Spider (Yu et al., 2019) database and table structure (currently focusing on single tables). Initially, we utilized GPT-4 to generate abstract data analysis questions relevant to real-world scenarios. Using these questions and the corresponding tables as instructions, we then employed GPT-4 to create diverse SQL statements along with their justifications. Following manual verification, we retained 400 samples to form the test dataset. The generated format is as follows:

```
[{
"sql": "data analysis SQL",
"title": "Data Analysis Title",
"showcase": "What type of charts",
"thoughts": "Current thinking",
},
...]
```

Examples of these tasks are in Appendix A.2.

BI Technical Skills Tasks

BI technical skills tasks primarily assess the ability of LLMs to not only understand BI knowledge but also simulate professional data analyst to apply the knowledge in solving realistic data analysis tasks. In task design, we extensively evaluate the model's various reasoning abilities, including two BI content inference tasks: utilizing SQL for solving specific data analysis problems and conducting in-depth analysis and summarization of data.

Text2SQL (3-1):*Convert natural language into SQL.* We sampled 600 entries categorized by industry type from the BIRDSQL (Li et al., 2023a) dataset.

Exploratory Data Analysis (3-2): *Provide data preprocessing suggestions through generated analysis code to streamline the data preparation process and ensure the quality of analysis.* We ob-

[5] www.biendata.xyz/competition/ccks_2020_4_1
[6] www.biendata.xyz/competition/ccks_2020_3
[7] aistudio.baidu.com/competition/detail/65
[8] https://github.com/MetaGLM/FinGLM

tained data analysis prompts and their corresponding code from the Kaggle website [9] through web scraping. By evaluating the combined count of stars and copies, we identified the code with the highest score as the answer to each prompt. After manual labeling and inspection, we randomly selected 100 samples as the test dataset.

Examples of these tasks are in Appendix A.3.

4 BIChat

4.1 Fine-tuing Data

To enhance our data analysis capabilities, we have constructed a domain-specific dataset comprising 400,000 examples by employing the Chain of Thought (COT) (Wei et al., 2022) and Self-Instruction (Wang et al., 2022) methods for data generation. This dataset includes instances for tasks such as NL2ViSQL, Doc2Markdown, Data2Insight, and Exploratory Data Analysis. The distribution of the dataset is illustrated in Figure 2.

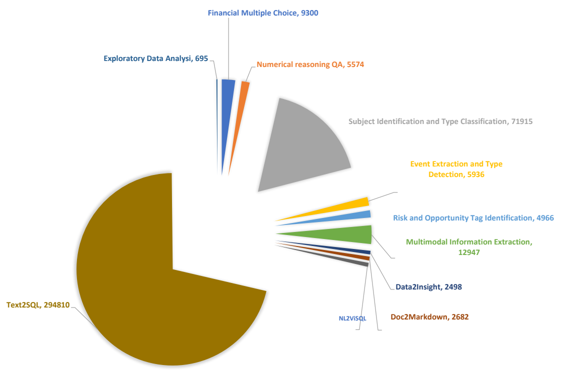

Figure 2: Data Distribution Across 11 Subtasks.

In addition, to facilitate more natural human-computer interactions, we amassed a significant collection of general-purpose dialogue datasets from major open-source repositories, including Alpaca (Taori et al., 2023), BELLE (Ji et al., 2023), MOSS (Sun et al., 2023), LIMA (Zhou et al., 2023a), among others. These datasets cover a wide variety of conversational contexts to ensure the model's capability for sustaining multi-turn interactions. The total dataset size amounts to approximately 1.89 million entries.

4.2 Data Preprocessing

To enhance data quality, we applied Dataset Quantization(DQ). (Zhou et al., 2023b) method to minimize similar data and increase the diversity of the dataset. Specifically, DQ first divides the entire dataset into a set of non-overlapping bins recursively based on the submodular gains that aims to maximize the diversity gains. Then, a small portion of data samples is uniformly sampled from all bins. In this manner, the selected samples are optimized to cover as much as possible the entire dataset with the interdata diversity maximized.

[9]www.kaggle.com/datasets?search=data+analysis

4.3 Training Approach of BIChat

Leveraging the Deepspeed (Yao et al., 2023) distributed training framework, we conducted training on a Qwen-7B-Chat (Bai et al., 2023) LLM using the QLoRA-based (Dettmers et al., 2023) approach across 8*RTX3090 GPUs with a total of 192GB of VRAM. During the training process, we incorporated LongLoRA (Chen et al., 2023b) and RoPE NTK (Liu et al., 2023b) interpolation methods to extend the fine-tuning sequence length to 6k tokens. Additionally, we substituted the traditional self-attention mechanism with Flash-attention2 (Dao, 2023), which significantly reduces memory usage and accelerates training speed compared to the original implementation.

Our training parameters were set as follows: QuantizationBit = 4, LoRARank = 64, LoRAAlpha = 128, LoRADropout = 0.05, LoRATarget = c_atten, MaxSeqLen = 2300 and Epochs=3.

5 Experiment

5.1 Evaluation Metrics

We defined 3 different metrics in total to measure different types of tasks:

Accuracy: Accuracy is a binary score that performs exact match between the model prediction and the gold answer. This applies to all single-label tasks including task 1-2.

F1: When there are multiple output labels, F1 score measures the harmonic mean of the precision and recall. This applies to all multi-label classification tasks including task 1-1, 2-1, 2-2, 2-3 and 3-4. In multiple-choice questions, answers that are selected too frequently (over-selected) or not selected enough (under-selected) are deemed incorrect.

Rouge-L: For other generation tasks 2-5, 2-6, 2-7, 3-1 and 3-2, we use the Rouge-L score to evaluate them. Rouge-L, commonly used in evaluation of generation tasks, automatically identifies the longest co-occurring n-gram sequences to compare the structural similarity of extracted answers with standard answers (Lin, 2004).The formula is ex-

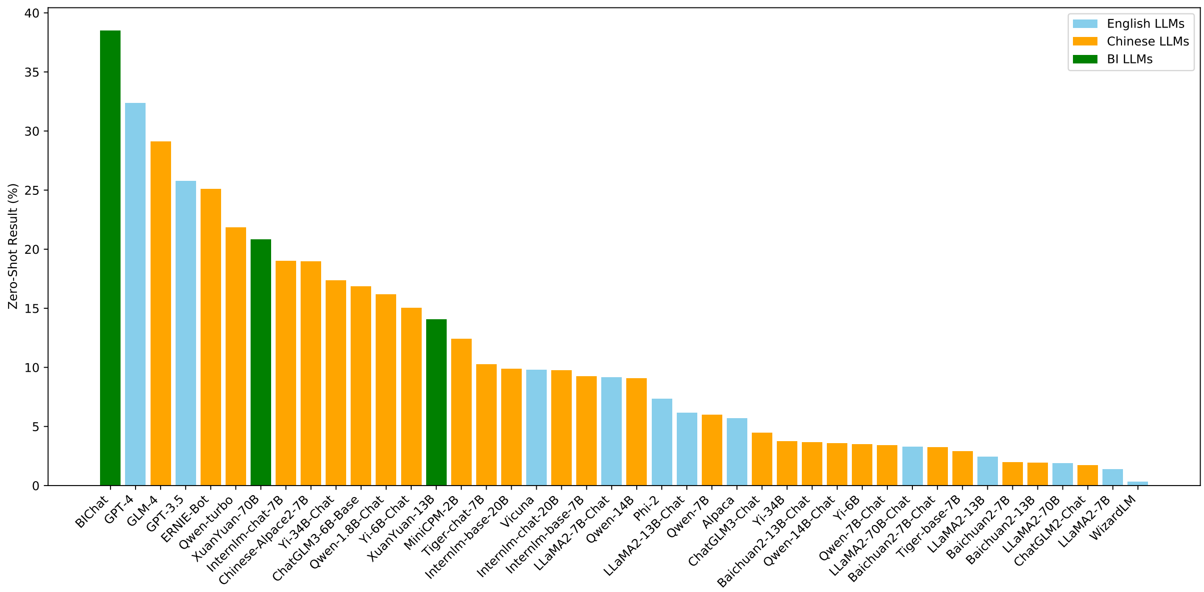

Figure 3: Average performance (zero-shot) of 42 LLMs evaluated on BIBench

pressed as follows, while β representing the weight between Precision and Recall.

$$F_{lcs} = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\text{Recall} + \beta^2 \cdot \text{Precision}} \quad (1)$$

5.2 Evaluated Models

We evaluate a wide spectrum of large language models of various sizes, grouping them into three major categories based on their pre-training and fine-tuning domains: multilingual LLMs, Chinese LLMs and BI LLMs. We provide a short review over them in the following section. The detailed model list is shown in Table 2.

English LLMs We consider 9 open-source english models: LLaMA-2-7B / 13B / 70B, LLaMA-2-Chat-7B / 13B / 70B, Alpacav1.0- 7B, Vicuna-v1.3-7B / 13B / 33B, WizardLM-7B. In addition, two commercial models, GPT-3.5-turbo-0613 and GPT-4-0613, are included. **Chinese LLMs** A number of Chinese LLMs have been proposed to enhance Chinese comprehension. They typically perform better than multilingual models on Chinese NLP tasks. We include 24 open-sourced, Chinese LLMs in our evaluation: Yi-Base 6B/34B, Yi-Chat 6B/34B, InternLM-Base 7B/20B, InternLM-Chat 7B/20B, Qwen-Base 7B/14B, Qwen-Chat 7B/14B, Baichuan2-Base 7B/13B, Baichuan2-Chat 7B/13B, TigerBot-Base-7B, TigerBot-Chat-7B, Chinese-Alpace2-7B, ChatGLM-6B, ChatGLM2-6B, ChatGLM3-Base-6B,ChatGLM3-6B. Moreover, three commercial models, 通义千问, 文心一言 and 智谱清言, are included. **BI LLMs** Currently, there is a lack of Chinese LLMs that have undergone additional fine-tuning with a Chinese corpus in the BIdomain to improve data analysis comprehension. As a result, we have turned our attention to LLMs from related fields. Here, we offer descriptions of the model: **XuanYuan-Chat:** based on LLaMA2-13B/70B, fine-tuned with general and financel instructions.

5.3 Experiment Setting

In the commercial models, we set the temperature to 0.7 and top p to 1. In other chat models, we tailor the prompt by using specific prefixes and suffixes for each model. Greedy decoding is performed during the generation process for all open-source models. We set the token length limit to 2400. Right truncation is performed for input prompts exceeding the length limitation. We evaluate all models in zero-shot settings, where the input for zero-shot inference consists solely of the task instruction.

5.4 Main Results

Figure 3 displays the overall zero-shot performance of each model. BIChat and GPT-4 are significantly ahead in the benchmarks, vastly outperforming all other models. With the same model size, LLMs that underwent Simplified Fine-Tuning (SFT) in Chinese outshine both the base Chinese LLMs and English SFT LLMs, demonstrating the effectiveness of fine-tuning on Chinese data. Furthermore, recent smaller models, like MiniCPM-2B, also exceed the

Model	Base Score	SFT Score	Diff Score
Tiger-7B	2.89%	10.27%	7.38%
Internlm-7B	9.22%	19.00%	9.78%
Baichuan2-13B	1.92%	3.67%	1.75%
Yi-34B	3.73%	17.34%	13.61%

Table 3: avg Performance score comparison of open source LLMs before and after SFT.

Task Name	LLaMA2 13B	XuanYuan 13B	LLaMA2 70B	XuanYuan 70B	Qwen-7B Chat	BIChat
1-1	18.80%	20.40%	9.60%	31.40%	4.40%	33.20%
1-2	0.31%	2.05%	0.00%	2.23%	0.00%	6.55%
2-1	0.42%	13.97%	0.43%	21.33%	3.47%	69.21%
2-2	0.01%	15.08%	1.10%	30.94%	10.98%	70.91%
2-3	0.11%	14.39%	0.98%	22.40%	0.60%	81.47%
2-4	0.28%	14.37%	1.52%	21.90%	9.66%	79.39%
2-5	0.39%	0.85%	1.38%	0.61%	0.30%	22.03%
2-6	0.17%	25.17%	4.59%	36.28%	1.28%	4.91%
2-7	1.74%	14.08%	0.31%	13.12%	0.50%	16.28%
3-1	1.69%	28.66%	2.23%	42.53%	5.18%	35.94%
3-2	0.38%	5.63%	1.23%	6.30%	1.06%	3.70%

Table 4: Comparison between different BI specific LLMs and their base models.

performance of many larger LLMs, indicating that the relationship between an LLM's capabilities and its size is not linear. Lastly, BI LLMs surpass general LLMs, suggesting that domain-specific fine-tuning can enhance a model's domain capabilities.

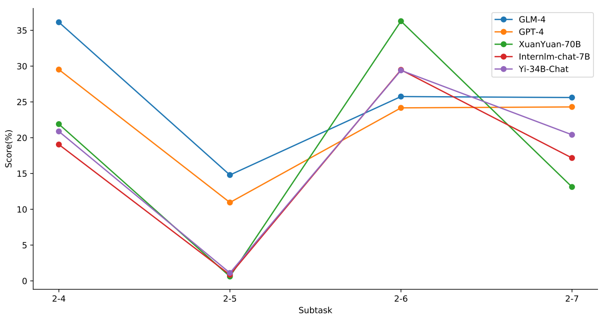

Figure 4: Comparison between different models for 2-4, 2-5, 2-6 and 2-7 tasks.

5.5 In-depth Analysis

Given the constraints on content, we have selected representative LLMs for in-depth analysis based on their types and high scores.

SFT may enhance model performance. As Table 3 demonstrates the open-source models' SFT versions outperform their Base counterparts. Notably, SFT data, collected from general domains, significantly improved model performance in BI tasks. Across models of equal size and architecture, performance variations suggest the training data's scope and specificity impact downstream tasks.

Most LLMs lack the capability for data analysis and insight generation. Sections 2-4 and 2-6 aim to assess LLMs' information parsing and processing abilities, while 2-5 and 2-7 focus on deriving insights and analyses based on data, evaluating whether large models can effectively extract viewpoints and information from structured or sequential data. We compared five different types of models and visualized the results of these four subtasks in Figure 4. The findings indicate that most models, including GPT-4 and GLM-4, show a comprehensive decline in performance, lacking data thinking and analytical abilities. This suggests that the challenge of developing LLMs capable of effectively generating insightful information remains an open problem.

BI-specific fine-tuning proves beneficial. To assess the impact of BI domain knowledge fine-tuning, we compared three LLMs, specifically fine-tuned with BI domain knowledge, against their corresponding base models, as shown in Figure 4. Notably, the XuanYuan and BIChat models demonstrate continuous score improvements following BI-specific knowledge fine-tuning. A closer examination of the 11 sub-tasks reveals that LLaMA2-13B, 70B, and Qwen-7B perform poorly across all tasks, indicating a lack of pre-training on a large-scale, high-quality data analysis corpus. Nonetheless, fine-tuning them with BI knowledge results in significant improvements. However, the models do not excel in tasks 3-2 and 2-5 post-fine-tuning, suggesting that fine-tuning alone may not suffice for complex BI tasks, possibly necessitating further research with Agents (Pan et al., 2024).

6 Conclusion

In this work, we presented the **BIBench**, an evaluation benchmark for assessing the capabilities of large language models in the field of data analysis field, comprising 11 tasks categorized across three cognitive dimensions. We undertake a thorough examination of 41 LLMs and assess their performance. The results demonstrate that current LLMs are still unable to give meaningful data analysis, and their scores on most tasks are often poor. While fine-tuning open-source LLMs(**BIChat**) on data analysis results in some advances, they still lag far below GPT-4 in some subtasks.

Limitations

Some of BIBench's data are sourced from the internet. Since existing large models often train on extensive internet data, there's a possibility that these models have already encountered a portion of the test data during their training. We intend to investigate more effective methods to prevent data pollution. Meanwhile, BIChat is trained using a 7B model, which may exhibit certain limitations in data analysis scenarios. Particularly in cases requiring fine-grained data insights, the model might not provide adequate data sensitivity and logical capability. To better adapt to data analysis scenarios, we plan to train BIChat using models with a parameter count of 14B or more.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023a. Felm: Benchmarking factuality evaluation of large language models. *arXiv preprint arXiv:2310.00741*.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Longlora: Efficient fine-tuning of long-context large language models. *arXiv:2309.12307*.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *Proceedings of EMNLP 2022*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

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)*, pages 320–335.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023a. Fingpt: Democratizing internet-scale data for financial large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. 2023b. Scaling laws of rope-based extrapolation.

Dakuan Lu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, Hengkui Wu, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed on December 28, 2023.

OpenAI. 2023. Gpt-4 technical report.

Haojie Pan, Zepeng Zhai, Hao Yuan, Yaojia Lv, Ruiji Fu, Ming Liu, Zhongyuan Wang, and Bing Qin. 2024. Kwaiagents: Generalized information-seeking agent system with large language models.

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. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2019. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task.

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

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xin Guo, and Yun Chen. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023a. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. 2023b. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17205–17216.

A Details of Task Instruction

A.1 BI Knowledge Memorization Tasks

INSTRUCTION: 运用你的相关知识回答以下问题

QUERY: 出口企业未将不同税率的出口货物分开报关、核算的，应从低适用退税率计算增值税出口退税。（）A:正确; B:错误

ANSWER: A; 分析解释：出口企业应将不同税率的货物分开核算和申报，未分开报关、核算或划分不清的，一律从低适用退税率计算退免税。

Table 5: The instruction and an example of Task 1-1 Financial Multiple Choice.

HISTORY:
26 | 2009 annual report in fiscal 2008 , revenues in the credit union systems and services business segment increased 14% (14%) from fiscal 2007 . all revenue components within the segment experienced growth during fiscal 2008 . license revenue generated the largest dollar growth in revenue as episys ae , our flagship core processing system aimed at larger credit unions , experienced strong sales throughout the year . support and service revenue , which is the largest component of total revenues for the credit union segment , experienced 34 percent growth in eft support and 10 percent growth in in-house support.
1	2008	year ended June 30 2009 2008	year ended June 30 2009 2008	year ended June 30 2009
—	——	————————	————————	——————
2	net income	$103102	$104222	$104681
3	non-cash expenses	74397	70420	56348
4	change in receivables	21214	-2913 (2913)	-28853 (28853)
5	change in deferred revenue	21943	5100	24576
6	change in other assets and liabilities	-14068 (14068)	4172	17495
7	net cash from operating activities	$206588	$181001	$174247
year ended june 30 , cash provided by operations increased $ 25587 to $ 206588 for the fiscal year ended june 30 , 2009 as compared to $ 181001 for the fiscal year ended june 30 , 2008 . this increase is primarily attributable to a decrease in receivables compared to the same period a year ago of $ 21214 . this decrease is largely the result of fiscal 2010 annual software maintenance billings being provided to customers earlier than in the prior year , which allowed more cash to be collected before the end of the fiscal year than in previous years .

INSTRUCTION: Provide numerical information related to user questions from tables and text, ensuring accurate calculation of results.

QUERY: what proportion does this represent?

ANSWER: 计算思考：第1步：add(2530454, 5923147) = 8453601.0；最终公式为：divide(5923147, add(2530454, 5923147))通过套用前面得到的数值，利用公式计算得到最终答案为：0.70067

Table 6: The instruction and an example of Task 1-2 Numerical reasoning QA.

A.2 BI Knowledge Understanding Tasks

820

INSTRUCTION: 定义了新闻舆情事件类型有：['业绩下滑', '提现困难', '交易违规', '失联跑路', '涉嫌违法', '不能履职', '涉嫌传销', '投诉维权', '财务造假', '涉嫌非法集资', '资金账户风险', '资产负面', '实控人股东变更', '高管负面', '信批违规', '评级调整', '涉嫌欺诈', '歇业停业', '重组失败', '履行连带担保责任', '债务违约', '业务资产重组', '股票转让-股权受让', '实际控制人变更', '债务重组', '商业信息泄露', '资金紧张', '实际控制人涉诉仲裁', '财务信息造假', '无']	
QUERY: 帮我判断新闻的舆情类型和对应的公司名称: "LG空调亏损严重或效仿新科两大缺陷遭退市尴尬华兰生物(002007)三季度净利下降45% 汇添富或为"失血门""跑路主力" 最后请以JSON的格式输出,格式参考如下："ps_type":"舆情类型", "company_name": "事件类型"　如果判断舆情类型为：无，则不需要抽取公司名称，则公司名称为无，格式为："ps_type":"无", "company_name": "无"　请不要返回与json内容无法的信息。	
ANSWER: {"ps_type": "业绩下滑", "company_name": "华兰生物"}	

Table 7: The instruction and an example of Task 2-1 Sentiment Analysis.

INSTRUCTION: 请以JSON List的格式输出,格式参考如下："type": ["事件类型", "事件类型"]请不要返回与json内容无法的信息。	
QUERY: 请帮我判断以下内容的事件类型: "在亿利洁能收购消息公布之后,其股价在数日之内大幅下跌,仅仅5个交易日时间股价便大跌38%。" 事件类型有：['质押', '投资', '股份股权转让', '减持', '起诉', '收购', '判决']	
ANSWER: {"type": ["收购"]}	

Table 8: The instruction and an example of Task 2-2 Event Extraction.

INSTRUCTION: None	
QUERY: 任务目标抽取出段落中主体公司和对应的机会标签和风险标签，段落内容：收购完成后，广州证券成为中信证券的全资子公司，并于**2020**年**3**月更名为中信证券华南公司，越秀金控则成为中信证券的第二大股东。标签列表如下：机会标签：'市场机遇': ['开户', '行业龙头', '中标', '注册', '星火', '业绩增加', '银行间市场', '金牛奖', '银团', '拿地', '重大事件利好', '获奖', '要约', '合作', '签订协议', '签约'], '政策机遇': ['建设基金', '批准成立', 'PPP', '批准通过', '批复', '批准筹建', '民营企业', '批准授权', '批准发行', '获批许可证', '批准进入', '政府引导基金', '资质证书'], '战略机遇': ['总部基地', '赎回票据', '自然人独资', '委托贷款', '兑换票据', '私有化', '股票回购', '现金管理', '履行程序', '闲置资金', '转让票据', '交债转债', '国债逆回购', '国企混改', '员工持股', '股东大会', '质押解除', '限制性股票', '股权奖励', '债券质押', '网银', '并购重组', '公司收购', '到期赎回', '票据合并', '股份出售']. 风险标签：'财务风险': ['负债', '收入预警', '破产清算', '财务异常', '财务风险', '兑付风险', '终止挂牌', '资产流失', '经济损失', '业绩下滑', '质押', '债务逾期', '资金短缺', '债务危机', '重大损失', '票据风险', '退市风险', '违约风险', '欠息', '信贷风险', '资产出售', '清盘'], '法律风险': ['拖欠工资', '司法拍卖', '被约谈', '监管处罚', '产品涉假', '停牌彻查', '进场核查', '资产冻结', '欺诈', '税务问题', '限制消费', '监管风险', '拖欠费用', '资金占用', '黑名单', '跑路', '查封', '吊销许可证', '经济纠纷', '侵权行为', '涉刑', '诉讼', '贪污涉贿', '违法违规', '专利纠纷'], '投融资风险': ['股改异常', '评级下降', '收购风险', '平仓风险', 'IPO遇阻', '注资异常', '股份减持', '发债遇阻', '民间融资', '壳资源', '治理风险': ['环保问题', '人事变动', '管理问题', '曝出', '质量事故', '安全隐患', '内部矛盾', '混乱'], '外部风险': ['安全生产事故', '意外事故', '工程受阻', '指责投诉', '舆论风险', '陷入局面', '工程事故', '投标受阻', '黑天鹅'] 标签类型可能有为空，一个或多个。最后请以List JSON的格式输出,格式参考如下："subject_company": "主体公司名称", "op_label":["机会标签",...], "risk_label":["风险标签", ...]。如果判断标签为空，需要返回主体公司即可，格式为："subject_company": 主体公司名称, "op_label":[], "risk_label":[]。请注意不要返回与json内容无法的信息。	
ANSWER: {"subject_company": "广州证券", "op_label": [], "risk_label": ["股权变动"]}	

Table 9: The instruction and an example of Task 2-3 Early Warning Analysis.

INSTRUCTION: None
QUERY: 定义了金融事件类型有：['公司上市', '股东减持', '股东增持', '企业收购', '企业融资', '股份回购', '质押', '解除质押', '企业破产', '亏损', '被约谈', '中标', '高管变动', '无']。请帮我判断以下内容的事件类型，事件类型可能有一个或多个，并事件类型抽取相应的触发词: "Verizon宣布62.5亿美元收购墨西哥美洲电信旗下无线服务提供商14日消息，美国移动运营商Verizon周一表示，将以62.5亿美元的现金加股票收购墨西哥美洲电信（America Movil）旗下的无线服务提供商Tracfone。此次收购预计将于2021年下半年完成，目前还在等待监管部门的批准。该笔交易将包括31.25亿美元的现金和31.25亿美元的Verizon普通股。它还包括与实现某些业绩衡量标准相关的最多额外6.5亿美元的未来现金对价。" 最后请以JSON List的格式输出,格式参考如下：["event_type": "事件类型", "trigger":"触发词", "event_type": "事件类型", "trigger":"触发词"]。请不要返回与json内容无法的信息。
ANSWER: [{"event_type": "企业收购", "trigger": "收购"}]

Table 10: The instruction and an example of Task 2-4 Multimodal Infor Extraction .

INSTRUCTION: 你是分析和挖掘数据中隐藏的深层信息和规律的专家，请用中文从不同的角度用3个要点生成一个专业的数据分析方案。
QUERY: 我向你提供了一个问题和这个问题的答案数据，请从数据分析常用的指标角度进行分析，给出计算后的指标结果，并结合你的知识给出针对性的相关建议。要求：不能超过2048个token，不要输出非分析性的内容。回答的格式如下: "1.xxxx","2.xxxx","3.xxxx"。问题："展示每位教员的id以及他/她指导的学生人数在散点图中"。答案数据："x_name": "FacID","y_name": "count(*)","x_data": [[1121, 1148, 2192, 2311, 5718, 7134, 7271, 7712, 7723, 7792, 8423, 8721, 8722, 8723, 8741, 8772, 8918, 9172]],"y_data": [[3, 3, 4, 3, 2, 2, 2, 1, 1, 1, 1, 1, 3, 1, 1, 3, 1, 1]]
ANSWER: "1.平均指导学生人数：通过计算y_data中学生人数的平均值，可以得到教员平均指导的学生人数。计算公式为：平均指导学生人数= 总学生人数/ 教员人数。由提供的数据计算得出，平均指导学生人数= (3+3+4+3+2+2+2+1+1+1+1+1+3+1+1+3+1+1) / 18 = 2。这表明教员平均每人指导2名学生。","2.指导学生人数分布情况：通过对y_data中的学生人数分布进行统计分析，我们可以了解教员指导学生数的集中趋势和离散情况。从数据中可以看出，大多数教员指导的学生人数集中在1-2人，少数教员指导的学生人数达到3-4人，很少有教员指导的学生人数超过4人。这可能意味着教学资源在教员之间分布不均。","3.教员指导学生数的离群值检测：在散点图中，可以识别离群值，即那些与大多数数据点显著不同的点。在此数据集中，教员指导学生人数的平均值为2人，如果某位教员指导的学生人数远高于或远低于这个平均值，那么这可以被认为是一个离群值。根据提供的数据，指导学生人数为4的教员可能是潜在的离群值，需要进一步分析他/她的指导工作是否高效，或者是否有过多的工作负担。"结合以上分析，建议学校管理部门：(1)平衡教员间的教学负担，以确保所有学生都能获得相同水平的指导和关注；(2)对那些指导学生人数超出平均值的教员进行进一步评估，确定是否需要调整其工作量或提供更多支持；(3)定期监控教员指导学生的数量，并在发现任何异常分布时采取适当的措施。

Table 11: The instruction and an example of Task 2-5 Data2Insigt.

INSTRUCTION: 如果给定的文本可以转换为Markdown语言的表格形式，其中包含上市公司年度报告的数据，并且能够准确反映各个数据之间的逻辑关系（例如，从A值到B值的盈利情况），那么将其转化为表格。如果无法转换为表格形式，则直接返回不支持。请注意，返回的内容要么是表格，要么是不支持，不会包含多余的内容。文本以换行符分割。
QUERY: 报告期内公司在加强内部生产管理和客户开拓的同时,也大幅增加了研发力度。2019年公司研发费用总额7,331万元,同比增长61.82%。截止到本报告期末,公司共有在研项目23项,主要研发方向包括新能源汽车用铝合金复合材料研发、多金属复合材料复合机理及工艺研发、大规格高强铝合金铸造机复合技术研发等,部分项目研发成果已经开始批量应用并获得了客户的广泛认可。
ANSWER: | 研发费用总额（万元）| 同比增长| |——————-|———| | 7,331 | 61.82% |

Table 12: The instruction and an example of Task 2-6 Doc2Markdown.

INSTRUCTION: According to the table schema and analyze goal, generate several SQL queries whose results can be view in one of the following chart type:['Table', 'LineChart', 'BarChart', 'IndicatorValue'] and respond in the following json format:"["sql": "data analysis SQL", "title": "Data Analysis Title", "showcase": "What type of charts to show", "thoughts": "Current thinking and value of data analysis"]
QUERY: 确定最大毛重最高的飞机。Table schema:['Aircraft_ID', 'Aircraft', 'Description', 'Max_Gross_Weight', 'Total_disk_area', 'Max_disk_Loading']
ANSWER: [{"sql": "SELECT Aircraft_ID, Aircraft, Description, Max_Gross_Weight FROM Aircraft ORDER BY Max_Gross_Weight DESC LIMIT 1", "title": "Aircraft with Maximum Gross Weight", "showcase": "Table", "thoughts": "通过查询飞机的最大起飞重量来找出最大的飞机。为了实现这个目标，我们选择从飞机数据库中选择飞机ID，飞机名称，描述和最大起飞重量字段，并按照最大起飞重量进行降序排序，然后只选择最大重量的那一架飞机。这个分析通过直接展示表格结果形式最为直观。"}, {"sql": "SELECT AVG(Max_Gross_Weight), AVG(Total_disk_area), AVG(Max_disk_Loading) FROM Aircraft", "title": "Statistics of key features", "showcase": "Table", "thoughts": "通过计算各航空器的总平均最大载重，总磁盘面积和最大磁盘载荷，我们可以得到重要特征的平均统计数据。这对于理解整体航空器性能具有重要指导价值。此分析结果适合表格形式展示，利于读者快速获取关键数据。"}]

Table 13: The instruction and an example of Task 2-7 NL2ViSQL.

A.3 BI Knowledge Applying Tasks

INSTRUCTION: 你是一个精通MySQl语法的SQL工程师，可以根据数据库的表结构和实际问题写出准确的、可以直接运行的MySQL查询语句。

QUERY: 回答结果为一个符合语法的SQL语句。在理解下面给出的数据库结构中各个数据表和数据项的含义后，生成解决问题的SQL语句（SQL语句尽可能简洁）:说出1945年上映的电影名称。按电影受欢迎程度的降序对列表进行排序。回答的格式为：xxxxx（SQL语句）该查询将在具有以下结构的数据库上运行: 数据表:数据表:'lists': '数据项': ['user_id', 'list_id', 'list_title', 'list_movie_number', 'list_update_timestamp_utc', 'list_creation_timestamp_utc', 'list_followers', 'list_url', 'list_comments', 'list_description', 'list_cover_image_url', 'list_first_image_url', 'list_second_image_url', 'list_third_image_url'], 'movies': '数据项': ['movie_id', 'movie_title', 'movie_release_year', 'movie_url', 'movie_title_language', 'movie_popularity', 'movie_image_url', 'director_id', 'director_name', 'director_url'], 'ratings_users': '数据项': ['user_id', 'rating_date_utc', 'user_trialist', 'user_subscriber', 'user_avatar_image_url', 'user_cover_image_url', 'user_eligible_for_trial', 'user_has_payment_method'], 'lists_users': '数据项': ['user_id', 'list_id', 'list_update_date_utc', 'list_creation_date_utc', 'user_trialist', 'user_subscriber', 'user_avatar_image_url', 'user_cover_image_url', 'user_eligible_for_trial', 'user_has_payment_method']

ANSWER:
```
SELECT movie_title
FROM movies
WHERE movie_release_year = 1945
ORDER BY movie_popularity DESC;
```

Table 14: The instruction and an example of Task 3-1 Text2SQL.

INSTRUCTION: Please help me analyze and summarize the task data.

QUERY: This dataset contains the historical data from 1960 to 2021 of the GDP by country, additionally its growth rate per year is calculated.The data is obtained from the World Bank data, the dataset is downloaded, a pre-processing was carried out in which geographic data such as regions, subregions were added and the % variation per year and country was calculated.The main objective of this dataset is to serve as a data source for the population analysis that I am developing to better understand the factors that affect population growth.

ANSWER:
```
# This Python 3 environment comes with many helpful analytics libraries installed
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for filename in filenames:
  print(os.path.join(dirname, filename))

# exploratory data analysis code
.....

# show figure
df.plot()
df.hist()
```

Table 15: The instruction and an example of Task 3-2 Exploratory Data Analysis.