Benchmarking Large Language Models on CFLUE - A Chinese Financial Language Understanding Evaluation Dataset

Jie Zhu[1], **Junhui Li**[2*], **Yalong Wen**[1], **Lifan Guo**[1]
[1]Alibaba Group, Hangzhou, China
[2]School of Computer Science and Technology, Soochow University, Suzhou, China
zhujie951121@gmail.com, lijunhui@suda.edu.cn
{wenyalong.wyl, lifan.lg}@alibaba-inc.com

Abstract

In light of recent breakthroughs in large language models (LLMs) that have revolutionized natural language processing (NLP), there is an urgent need for new benchmarks to keep pace with the fast development of LLMs. In this paper, we propose CFLUE, the Chinese Financial Language Understanding Evaluation benchmark, designed to assess the capability of LLMs across various dimensions. Specifically, CFLUE provides datasets tailored for both knowledge assessment and application assessment. In knowledge assessment, it consists of 38K+ multiple-choice questions with associated solution explanations. These questions serve dual purposes: answer prediction and question reasoning. In application assessment, CFLUE features 16K+ test instances across distinct groups of NLP tasks such as text classification, machine translation, relation extraction, reading comprehension, and text generation. Upon CFLUE, we conduct a thorough evaluation of representative LLMs. The results reveal that only Qwen-72B, GPT-4, and GPT-4-turbo achieve an accuracy exceeding 60% in answer prediction for knowledge assessment, suggesting that there is still substantial room for improvement in current LLMs. In application assessment, while GPT-4 and GPT-4-turbo rank as the top two performers on average, their significant advantage over open-source LLMs is noticeably diminished, given that Qwen-72B achieves the best performance in 2 out of 5 tasks. The datasets and scripts associated with CFLUE are openly accessible at https://github.com/aliyun/cflue.

1 Introduction

Recently, the remarkable capabilities exhibited by large language models (LLMs) have brought significant advancements and revolutionized natural language processing (NLP). In response to the emergence of LLMs, new benchmarks have been introduced to assess a varied range of abilities exhibited by these models. Notable examples include MMLU (Hendrycks et al., 2021), HELM (Liang et al., 2023), BIG-bench (Srivastava et al., 2023), and GLUE (Wang et al., 2019b), which are widely employed to evaluate LLMs' capabilities in English. Additionally, benchmarks such as CLEU (Xu et al., 2020), CMMLU (Li et al., 2023), and C-Eval (Huang et al., 2023) are widely used to evaluate their capabilities in Chinese.

However, assessing LLMs in the financial domain, especially within the Chinese context, presents considerable challenges due to limitations in existing Chinese financial evaluation datasets. Notably, datasets like FinanceIQ (Duxiaoman-DI, 2022) and FinEval (Zhang et al., 2023) have restrictions in terms of size and diversity, hindering a comprehensive evaluation of LLM performance. While FinanceIQ and FinEval focus solely on multiple-choice tasks, there is a lack of datasets allowing a thorough examination of LLM capabilities in text generation. Additionally, existing shared tasks, such as those in CCKS (Tianchi, 2019, 2020, 2021, 2022), predominantly concentrate on event extraction tasks, limiting the objective and quantitative measurement of LLM performance.

Inspired by the FLUE benchmark (Shah et al., 2022), which encompasses a comprehensive set of datasets across five financial domain tasks in English, this paper introduces a novel dataset named CFLUE (Chinese Financial Language Understanding Evaluation). CFLUE addresses the aforementioned challenges by providing a benchmark for evaluating LLM performance through various NLP tasks, categorized into knowledge assessment and application assessment. In the knowledge assessment section, CFLUE contains 38K+ questions from 15 types of financial qualification mock exams of varying difficulty levels and subjects. These questions are in the multiple-choice question format to enable standardized and objective evalua-

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2024, pages 5673–5693

tions. Additionally, each question is accompanied by a solution explanation, facilitating open-ended or chain-of-thought assessments of LLMs' reasoning capabilities. In the application assessment section, CFLUE contains 16K+ test instances covering five groups of classical NLP tasks: text classification, machine translation, relation extraction, reading comprehension, and text generation. These instances are sourced from existing shared tasks or are annotated by professionals using real data sources.

Based on CFLUE, we assess the effectiveness of several representative LLMs across general and financial domains. This includes an examination of three OpenAI LLMs, nine open-source LLMs in the general domain, and three open-source LLMs in the financial domain. From the experimental results, we have the following key findings:

- The findings are not consistent in terms of tasks in knowledge assessment and application assessment. In knowledge assessment, GPT-4 and GPT-4-turbo (OpenAI, 2023) demonstrate significant superiority over other LLMs, achieving over 60% accuracy in answer prediction. However, this suggests substantial room for improvement in current LLMs. In application assessment, although GPT-4 and GPT-4-turbo remain top performers on average, and in specific tasks, they may even fall behind certain open-source LLMs. This could be attributed to the specialized design of these open-source LLMs, which are tailored for Chinese data.

- Existing financial domain LLMs, such as FinGPT (Yang et al., 2023), DISC-FinLLM (Chen et al., 2023), and Tongyi-Finance (TongyiFinance, 2023), exhibit poor zero-shot performance in both knowledge and application assessment tasks, indicating limited coverage of financial knowledge and significant room for improvement.

- Open-source LLMs benefit significantly from supervised fine-tuning. For instance, ChatGLM3-6B (Zeng et al., 2022), Qwen-7B (Bai et al., 2023), and Baichuan2-7B (Baichuan, 2023) achieve comparable or superior performance to ChatGPT (OpenAI, 2022) in both answer prediction and reasoning tasks, despite having only 4% of ChatGPT's parameters.

In summary, this paper provides valuable insights into the performance of LLMs in Chinese financial contexts from multiple perspectives. Our findings suggest that there still exists much room for improvement even for the current best performers. We hope that CFLUE could guide the developers to understand the abilities of their models from multiple dimensions and facilitate the growth of foundation models in Chinese financial domain.

2 Related Work

Financial Evaluation Datasets. A summary of recent benchmarks in the financial domain is presented in Table 1. In English, FINQA by Chen et al. (2021) introduces a dataset comprising 8,281 question-answering pairs, emphasizing numerical reasoning processes. Both TAT-QA by Zhu et al. (2021) build a large-scale QA dataset containing both Tabular And Textual data from real financial reports. BizBench by Koncel-Kedziorski et al. (2023) consists of 8 quantitative reasoning tasks of question-answering (QA) for structured and unstructured financial data via program synthesis. FINANCEBENCH by Islam et al. (2023) consists of 10,231 questions about publicly traded companies, with corresponding answers and evidence strings. FLUE by Shah et al. (2022) utilizes datasets from existing literature to establish comprehensive benchmark across five tasks in financial domain, including sentiment analysis classification from Financial PhraseBank dataset (Malo et al., 2014) and sentiment analysis regression from FiQA 2018 Task-1 Maia et al. (2018),[1] named entity recognition (NER) from NER data on loan agreement (Alvarado et al., 2015), question answering from FiQA 2018 Shared Task-2 Maia et al. (2018), news headline classification from gold commodity news (Sinha and Khandait, 2020), and structure boundary detection from FinSBD-3 Shared Task (FinSBD3, 2021).

In Chinese, the CCKS series since 2019 CCKS has released various datasets for event extraction tasks (Tianchi, 2019, 2020, 2021, 2022). Additionally, both FinanceIQ and FinEval offer thousands of multiple-choice question answering pairs that could be used as evaluation suites for LLMs (Duxiaoman-DI, 2022; Zhang et al., 2023). However, compared to English, Chinese datasets tend to focus on either event extraction (Tianchi, 2022) or multiple-choice question-answering (Duxiaoman-DI, 2022; Zhang et al., 2023; Yang et al., 2023), demonstrating limited task diversity. In contrast, our paper introduces

[1] https://sites.google.com/view/fiqa

Language	Source	Question Type (Task)
English	FINQA (Chen et al., 2021)	Question Answering
	TAT-QA (Zhu et al., 2021)	Question Answering
	BizBench (Koncel-Kedziorski et al., 2023)	Quantitative Reasoning
	FINANCEBENCH (Islam et al., 2023)	Question Answering
	FLUE (Shah et al., 2022)	Sentiment Classification (Malo et al., 2014) Sentiment Analysis (FiQA (FiQA, 2018)) News Headline Classification (Sinha and Khandait, 2020) Named Entity Recognition (Alvarado et al., 2015) Structure Boundary Detection (FinSBD3, 2021) Question Answering (FiQA (FiQA, 2018))
Chinese	CCKS (Tianchi, 2019, 2020, 2021, 2022)	Event Extraction, Event Entity (and Causality) Extraction
	FinanceIQ (Duxiaoman-DI, 2022)	Multiple-choice Question Answering
	FinEval (Zhang et al., 2023)	Multiple-choice Question Answering
	FinGPT-fineval (Yang et al., 2023)	Multiple-choice Question Answering
	CFLUE (Ours)	Multiple-choice Question Answering & Reasoning Text Classification Machine Translation Relation Extraction Reading comprehension Text Generation

Table 1: A summary of benchmarks in the financial domain.

CFLUE (Chinese Financial Language Understanding Evaluation), featuring a set of heterogeneous benchmark tasks for a more comprehensive evaluation. Additionally, beyond evaluation suites, there are financial datasets like SmoothNLP[2], IREE (Ren et al., 2022), suitable for training or fine-tuning models in the finance domain.

Other Benchmark Datasets. The development of LMs (Devlin et al., 2019; Radford et al., 2019) has witnessed heterogeneous benchmarks to probe their diverse abilities. In English, traditional benchmarks focus on single tasks such as natural language understanding (Wang et al., 2019b,a), reading comprehension (Rajpurkar et al., 2018; Dua et al., 2019), and reasoning (Zellers et al., 2019; Sakaguchi et al., 2021). Recently, researchers have introduced broader benchmarks like TruthfulQA (Lin et al., 2022), evaluating truthfulness across 38 categories spanning various domains. MMLU (Hendrycks et al., 2021) offers a multi-task evaluation covering 57 tasks across different fields. Additionally, BIG-bench (Srivastava et al., 2023) and HELM (Liang et al., 2023) benchmarks cover as much to 204 and 42 tasks, respectively.

In Chinese, CLEU (Xu et al., 2020) is the first large-scale Chinese benchmark that covers tasks such as classification and reading comprehension. AIGEval (Zhong et al., 2023) evaluates LLM capabilities in human-centric exams, including college entrance exams, law school admission tests, math competitions, and lawyer qualification tests. Both CMMLU (Li et al., 2023) and CEval (Huang et al., 2023) are comprehensive Chinese benchmarks spanning multiple levels and diverse subjects. CMExam (Liu et al., 2023) focus on the Chinese medical domain, providing multiple-choice questions from the Chinese National Medical Licensing Examination.

3 CFLUE: Chinese Financial Language Understanding Evaluation

While various LLMs may exhibit similar performance on tasks within the general domain, it is often the domain-specific tasks that set them apart. The primary objective of CFLUE is to assess the financial domain knowledge capabilities of models from multiple dimensions. As illustrated in Figure 1, CFLUE evaluates the proficiency of LLMs from both knowledge and application perspective. The statistics of CFLUE are presented in Table 2.

3.1 Knowledge Assessment

This assessment consists of questions presented in a multiple-choice format, sourced from real-world, challenging mock exams accessible through public channels. The primary origins include 15 types of qualification exams covering diverse difficulty levels and subjects. The mock exams are exclusively sourced from PDF or Microsoft Word documents available on the Internet, avoiding direct extraction from plain text webpages. Using tools like pdfplumber[3] and PaddleOCR[4], we have collected a

[2] https://github.com/smoothnlp/FinancialDatasets

[3] https://github.com/jsvine/pdfplumber

[4] https://github.com/PaddlePaddle/PaddleOCR

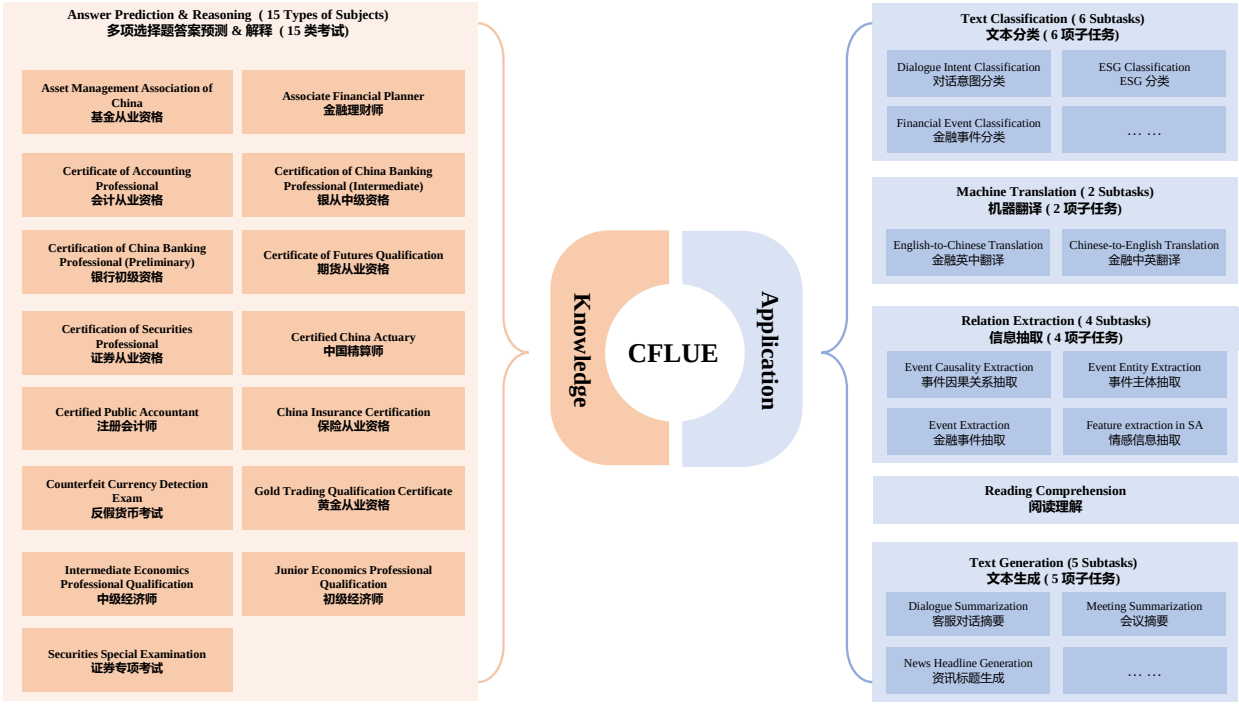

Figure 1: Overview diagram of CFLUE benchmark.

Type	Task	Size			Avg. Length	Description
		Train	Valid	Test		
Knowl.	Answer Prediction & Reasoning	30,908	3,864	3,864	130	15 types of qualification exams
Appl.	Text Classification	-	-	3,312	840	6 subtasks
	Machine Translation	-	-	3,000	51	2 translation directions
	Relation Extraction	-	-	3,500	274	4 subtasks
	Reading Comprehension	-	-	2,710	201	in question answering format
	Text Generation	-	-	4,000	947	5 subtasks
	Total	-	-	16,522	-	-

Table 2: Statistics of CFLUE. The average length indicates the average number of words in the inputs.

total of 166,681 multiple-choice questions.

We then exclude questions that relied on non-text information, such as those containing HTML tags, images, tables, and questions with keywords 图/image and 表/table. Additionally, we remove duplicate questions from the dataset. In an additional step to minimize the risk of data contamination, we employ GPT-4 to rephrase the questions and shuffle the multiple choices, we randomly shuffle the order of all options and use GPT-4 to rewrite the questions, enhancing the diversity of the dataset.

After preprocessing, we retain 38,636 questions. Among them, 60% of the questions have up to six options with only one correct answer, 10% are true/false questions, and the remaining questions may have multiple correct options. Each question in the dataset has a unique ID and is accompanied by detailed explanations provided by professionals to facilitate a thorough evaluation of the language model's logical reasoning abilities. Figure 2 in Appendix gives an example of annotated multi-choice question.

3.2 Application Assessment

This assessment consists of five sets of real-word financial NLP tasks (Section 3.2.1∼ 3.2.5). Then we discuss our strategy for quality control in Section 3.2.6.

3.2.1 Text Classification

It covers the following five subtasks.

- Dialogue intent classification. We randomly sample 500 instances from Banking dataset (Casanueva et al., 2020)[5] which consists of 13,083 instances across 77 intends. The sampled 500 instances are then translated into Chinese by GPT-4, followed by manual verification.

[5] https://github.com/PolyAI-LDN/task-specific-datasets

- ESG (environment, social and governance) classification and ESG sentiment classification. It contains 180 instances belonging to 14 classes for ESG classification and 3 classes for ESG sentiment classification. All instances are sourced from news portals such as 证券时报/Securities Times, 每日经济新闻/Daily Economic News, and 上海证券报/Shanghai Securities News, etc. Two senior doctoral students carry out the annotations following the CSMAR guideline available in https://data.csmar.com/.

- Financial Event Classification. We randomly select 1,000 instances from CCKS 2020,[6] distributed across 27 classes, with each class having a similar number of instances.

- Financial Industry Classification. It contains 1,000 instances belonging to 68 industries. All the instances are news from EastMoney website[7], each labeled with its corresponding industry.

- Financial Meeting Content Classification. It contains 452 instances distributed among three classes. The ASR toolkit[8] is first utilized to convert audio files from brokerage firms' meetings into text. Annotators then segment each meeting text and categorize segments as either 财务情况/Financial Situation, 未来展望/Future Outlook, or 经营情况/Operational Status. Segments not falling into these classes are excluded.

3.2.2 Machine Translation

We gather Chinese-English bilingual economic news reports from https://www.kekenet.com/course/13650/. A total of 1,500 sentence pairs, each containing more than 30 words on each side, are selected. These sentence pairs serve as the dataset for both Chinese-to-English and English-to-Chinese translation.

3.2.3 Relation Extraction

It covers the following four subtasks.

- Feature extraction in Sentiment Analysis. We collect 500 stock comments from a financial company, and annotators mark the subject of the comments along with their sentiment, covering positive, neutral, and negative.

- Financial Event Causality Extraction. We randomly select 1,000 instances from CCKS 2021 task6-2.[9]

- Financial Event Entity Extraction. We randomly select 1,000 instances from CCKS 2019.[10]

- Financial Event Extraction. We randomly select 1,000 instances from DuEE-fin (Han et al., 2022).[11]

3.2.4 Reading Comprehension

Initially, we set up a document repository comprising approximately 90,000 items, which includes news articles, research reports, and annual reports. Subsequently, we generate a list of questions by drawing from the common queries encountered in real customer scenarios, involving input from over 20 financial analysts. For each question, we employ a standard extraction framework[12] (utilizing es+m3e embedding[13]) to establish a RAG (Retrieval-Augmented Generation) process, resulting in the most relevant chunk referred to as the financial text. Following this, for each (financial text, question) pair, annotators evaluate and determine if the question can be answered based on the financial text. If it is feasible, analysts write down the answer based on the financial text. Conversely, if not, analysts mark the answer as 根据提供的背景信息，无法回答该问题/*According to the provided information, I cannot answer that question*. After removing similar questions, we retain 2,710 triples of (financial text, question, answer), with 245 questions (9%) that cannot be answered based on the financial text.

3.2.5 Text generation

It covers the following five subtasks.

- Dialogue Summarization. We randomly select 1000 instances from CSDS (Lin et al., 2021).[14] CSDS offers both overall summaries and role-oriented summaries. We utilize the overall summary for dialogue summarization.

- Meeting Summarization. We employ the identical audio files utilized in Financial Meeting Content Classification. Subsequently, we invite three

[6] https://tianchi.aliyun.com/dataset/111209
[7] https://www.eastmoney.com/
[8] https://github.com/alibaba-damo-academy/FunASR
[9] https://www.biendata.xyz/competition/ccks_2021_task6_2/
[10] https://www.biendata.xyz/competition/ccks_2019_4
[11] https://www.luge.ai/#/luge/dataDetail?id=7
[12] https://www.langchain.com/
[13] https://huggingface.co/moka-ai/m3e-base
[14] https://github.com/xiaolinAndy/CSDS

annotators to annotate the summary for each meeting.

- News Headline Generation. We use the same news articles as those utilized in Financial Industry Classification. Note that each news article is associated with a corresponding headline.

- Research Report Headline Generation. We retrieve research reports from the EastMoney website. Likewise, each research report is associated with a corresponding headline.

- Term Interpretation. We extract financial terminologies and their explanations from ChinaValue website.[15]

3.2.6 Quality Control

The annotated datasets are sourced from our diverse projects, with 47 experienced annotators engaged for CFLUE. Ranging from senior doctoral students to finance professionals like treasury analysts and MBAs, they undergo training and access guidance documents. Additional senior analysts for each project oversee their work, offering feedback.

- For both ESG Classification and ESG Sentiment Classification, two senior doctoral students are recruited. Over several weeks, tasks were gradually assigned as annotators gained proficiency. Weekly, each annotator tackled around 20-25 articles, with five shared between them. Discrepancies in annotations were resolved through discussion, and remaining articles were revisited afterward.

- For both Financial Meeting Content Classification and Meeting Summarization, 20 analysts are recruited to segment meeting content and then categorize and summarize each segment. Using ASR output, annotators initially segment the text based on the topic.[16] Only segments with well-defined boundaries are chosen for classification and summarization. Before annotation, comprehensive training is provided, and discussions involving 10 ASR texts ensure consensus. The project spans several weeks due to the time-consuming nature of both text segmentation and summarization.

- For Feature extraction in Sentiment Analysis, five analysts label the subject of the comments and their corresponding sentiment. Before annotation, they undergo training and discuss 20 comments until a high level of agreement is achieved. From the total 1,115 instances collected, 500 are randomly chosen based on length.

- For reading comprehension, 20 analysts gather queries and mark answers from related financial text. Before annotation, they undergo training, and discuss 10 pairs of (financial text, question) until a high level of agreement is achieved. Triples are discarded if the financial texts are not relevant to their corresponding questions. The project spans three months, with each annotator averaging 300 triples. Out of 6,000 triples, 2,710 are retained by removing similar questions.

3.3 Data Statistics

Table 2 shows the statistics of all CFLUE tasks after preprocessing. In the knowledge assessment's multiple-choice question answering, questions are randomly divided into training, validation, and test sets at a ratio of 8:1:1, with the validation set designated for hyperparameter tuning.[17] As for tasks in the application assessment, they exclusively comprise test questions, ranging from 2.7K to 4.0K. For detailed statistics on these datasets, please refer to Table 6 and Figure 3 in Appendix A.

4 Experimentation

4.1 Models, Settings, Prompts, and Metrics

Models. For a comprehensive assessment of the state of LLMs in the Chinese language context, we categorize the benchmarked LLMs into two groups: LLMs of the general domain and LLMs of the financial domain, as shown in Table 3.

- **LLMs of the general domain:** This group consists of models trained on extensive volume of general-purpose corpora. Included in this group are three OpenAI LLMs –GPT-4-turbo, GPT-4 (OpenAI, 2023), and ChatGPT (GPT-3.5-turbo) (OpenAI, 2022) –along with five families of nine open-source LLMs, LLaMA2-7B/-

[15] http://www.chinavalue.net/Wiki

[16] The segmentation is performed using the automatic segmentation result from the toolkit https://www.modelscope.cn/models/iic/nlp_bert_document-segmentation_chinese-base/summary

[17] We acknowledge that, in the context of few-shot and chain-of-thought evaluation, the questions in the knowledge assessment could be partitioned into dev, valid, and test sets, such as at a ratio of 1:1:8. However, in this paper, we opt not to adhere to this division due to budget constraints on OpenAI API calls.

Domain	Model	Creator	#Param.	Oriented	Access
General	GPT-4-turbo	OpenAI	N/A	English	API
	GPT-4	OpenAI	N/A	English	API
	ChatGPT	OpenAI	N/A	English	API
	LLaMA2-7B/-70B	Meta	7B/70B	English	Weights
	Vicuna v1.5-7B	Chiang et al. (2023)	7B	English	Weights
	ChatGLM3-6B	Tsinghua	6B	Chinese	Weights
	Qwen-7B/-14B/-72B	Alibaba	7B/14B/72B	Chinese	Weights
	Baichuan2-7B/-13B	Baichuan	7B/13B	Chinese	Weights
Financial	FinGPT-6B	Yang et al. (2023)	6B	Chinese	Weights
	DISC-FinLLM-13B	Fudan	13B	Chinese	Weights
	Tongyi-Finance-14B	Alibaba	14B	Chinese	Weights

Table 3: LLMs benchmarked in this paper.

70B (Touvron et al., 2023), Vicuna v1.5-7B (Chiang et al., 2023), ChatGLM3-6B (Zeng et al., 2022), Qwen-7B/-14B/-72B (Bai et al., 2023), and Baichuan2-7B/-13B (Baichuan, 2023).

- **LLMs in the financial domain:** This group comprises several representative LLMs specifically tailored for the financial domain. FinGPT-6B (Yang et al., 2023) (fingpt-mt_chatglm2-6b_lora) is initialized with ChatGLM2-6B parameters and fine-tuned using LoRA (Hu et al., 2022) on datasets covering multiple tasks, including financial sentiment analysis, relation extraction, headline classification, and named entity recognition. DISC-FinLLM-13B (Chen et al., 2023) is based upon Baichuan-Chat-13B, with all parameters fine-tuned on Chinese financial corpora. Similarly, Tongyi-Finance-14B (TongyiFinance, 2023) is based on Qwen-14B and fine-tuned on Chinese financial corpora as well.

Among the aforementioned LLMs, GPT-4 turbo, GPT-4, ChatGPT, LLaMA2, and Vicuna v1.5 are English-oriented, while the remaining models are tailored for Chinese.

Supervised Fine-Tuning Settings. For knowledge assessment, we conduct additional fine-tuning of the open-source models using the training dataset. Specifically, we employ LoRA (Hu et al., 2022) to fine-tune LLaMA2-7B, Vicuna v1.5-7B, ChatGLM3-6B, Qwen-7B, and Baichuan2-7B, with the rank set to 8, alpha to 32, and dropout to 0.1. During the fine-tuning, a single NVIDIA A100/80G GPU is utilized. The batch size is set to 16 with a gradient accumulation step of 8.

Prompts. Our evaluation of LLMs is conducted in a zero-shot setting.[18] In the knowledge assessment, we follow the approach of Liu et al. (2023), wherein we perform answer prediction and reasoning simultaneously by instructing LLMs to generate both the answer and a solution explanation. Specifically, we use three similar prompts, one for each question type. In the application assessment, we use a single prompt for each subtask, maintaining consistency by employing similar prompts for subtasks within each group. We present a few prompt examples in Appendix C.

Metrics. In the knowledge assessment, we measure the performance of answer prediction using accuracy and weighed F1 score. Additionally, for the open-ended reasoning we report BLEU (Papineni et al., 2002) and ROUGE (Lin and Hovy, 2002) scores. In the application assessment, we report accuracy for text classification, BLEU and COMET (Rei et al., 2020) for machine translation, F1 score for relation extraction, ROUGE for reading comprehension and text generation.

4.2 Results of Knowledge Assessment

Table 4 compares the performance of general domain LLMs and financial domain LLMs in the contexts of answer prediction and reasoning.[19] Regarding the answer prediction task, we have the following observations:

- GPT-4-turbo and GPT-4 exhibit superior performance compared to ChatGPT and other LLMs except Qwen-72B. These two models achieve a notable 60% accuracy and 0.60 in F1, highlighting the challenges posed by CFLUE. Consistent with observations in other Chinese benchmarks (Huang et al., 2023; Liu et al., 2023), GPT-4 substantially outperform ChatGPT with a performance gap of 17% difference in accuracy.

[18] This paper focuses on the zero-shot setting, given the well-defined nature of all tasks in our assessment. We leave the exploration of the few-shot setting in our future work.

[19] Table 7 and Table 8 in Appendix B offer a comprehensive breakdown of performance, detailing results for each subject and question type.

Domain	Model	Prediction		Reasoning				
		Acc (%)	F1 (%)	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
General	GPT-4-turbo	60.61±0.21	60.31±0.19	30.66±0.22	10.61±0.13	40.28±0.20	17.23±0.15	28.62±0.19
	GPT-4	60.87±0.11	60.82±0.10	37.58±0.18	17.26±0.09	44.50±0.12	22.42±0.08	32.59±0.11
	ChatGPT	43.35±0.60	42.96±0.70	41.67±0.76	20.46±0.51	47.37±0.19	25.29±0.18	35.41±0.13
	LLaMA2-7B	17.66±0.39	10.34±0.31	9.46±0.16	3.93±0.10	17.77±0.17	7.65±0.16	15.48±0.18
	LLaMA2-70B	18.79±0.25	15.54±0.21	13.11±0.11	5.49±0.07	22.02±0.19	9.72±0.14	19.06±0.2
	Vicuna v1.5-7B	31.14±0.37	30.92±0.35	29.6±0.21	12.92±0.16	40.68±0.11	19.32±0.11	34.27±0.07
	ChatGLM3-6B	40.78±0.33	41.37±0.33	34.7±0.47	16.74±0.23	43.74±0.08	22.92±0.09	37.68±0.04
	Qwen-7B	43.63±0.37	43.25±0.41	42.03±0.32	17.85±0.29	39.87±0.26	22.11±0.21	35.06±0.28
	Qwen-14B	53.82±0.23	54.23±0.27	40.05±0.34	21.56±0.25	47.61±0.11	27.27±0.10	41.45±0.12
	Qwen-72B	**72.8±0.23**	**73.04±0.23**	45.78±0.39	26.76±0.21	50.78±0.15	31.48±0.13	**45.28±0.15**
	Baichuan2-7B	32.31±0.14	28.77±0.19	21.71±1.36	0.17±0.08	7.54±0.12	3.23±0.09	6.9±0.12
	Baichuan2-13B	41.5±0.29	40.87±0.29	28.64±0.57	14.16±0.28	42.04±0.06	22.36±0.10	36.51±0.05
Financial	FinGPT-6B	34.27±0.36	35.88±0.22	28.28±0.61	13.55±0.23	39.08±0.16	19.92±0.13	33.58±0.14
	DISC-FinLLM-13B	35.25±0.28	33.79±0.24	29.66±0.21	1.69±0.15	14.92±0.12	6.56±0.18	11.20±0.13
	Tongyi-Finance-14B	47.21±0.11	47.07±0.16	38.32±0.11	19.24±0.05	44.35±0.07	23.55±0.06	38.1±0.10
	LLaMA2-7B-CFLUE	27.07±0.65	26.93±0.65	36.7±1.60	18.56±0.22	43.29±0.19	23.72±0.16	38.22±0.16
	Vicuna V1.5-7B-CFLUE	29.84±0.41	29.25±0.43	47.37±1.71	19.27±0.24	48.17±0.21	28.21±0.17	38.12±0.18
	ChatGLM3-6B-CFLUE	42.43±0.24	41.93±0.27	42.45±1.39	19.55±0.76	43.06±0.30	24.08±0.30	38.17±0.29
	Qwen-7B-CFLUE	48.61±0.58	48.59±0.6	**48.19±1.94**	**22.94±1.14**	47.62±0.19	27.73±0.17	42.41±0.15
	Baichuan2-7B-CFLUE	47.33±0.34	47.20±0.38	47.69±1.45	23.11±0.91	**53.02±0.41**	**33.19±0.39**	42.98±0.27
Random	Random	21.92±0.46	21.58±0.46	-	-	-	-	-

Table 4: Overall comparison on the answer prediction and reasoning tasks of knowledge assessment. Scores in **bold**/underline denote the top/second-best performances.

- Among the nine open-source general domain LLMs with sizes ranging from 6B to 72B parameters, Qwen-72B attains the best performance, followed by Qwen-14B and Qwen-7B which even surpass ChatGPT. The performance gap between the three Qwen models also suggests that increasing LLM size significantly benefits downstream tasks. ChatGLM3-6B and Baichuan2-13B achieve similar performance, slightly underperforming ChatGPT. Beichuan2-7B achieves 36% accuracy and 0.35 in F1, followed by Vicuna v1.5-7B, LLaMA2-70B and LLaMA2-7B.

- Among the three LLMs designed for the financial domain, Tongyi-Finance-14B attains the highest performance. Despite being built upon Qwen-14B, Tongyi Finance-14B unexpectedly lags behind with a 10% lower accuracy. In comparison to general domain LLMs, financial domain LLMs struggle to achieve satisfactory performance. This finding aligns with that of Liu et al. (2023). Similarly, financial domain LLMs may suffer from limited corpus diversity, hindering the acquisition of broad financial knowledge needed for CFLUE questions.

- Fine-tuning general domain LLMs with LoRA on the CFLUE dataset significantly boosts their performance. For example, the accuracy of Qwen-7B increases to 49.84% from 45.70%. Even with only 4% of the parameters used in ChatGPT, Qwen-7B-CFLUE, Baichuan2-7B-CFLUE, and ChatGLM3-6B-CFLUE all surpass ChatGPT.

In the reasoning task, the performance trend differs from the answer prediction task. Qwen-72B excels in BLEU and ROUGE metrics, outperforming other models. In contrast, ChatGPT surpasses GPT-4 and GPT-4-turbo, while ChatGLM3-6B and Qwen-7B match or surpass GPT-4. However, models like LLaMA2-7B, Baichuan2-7B, and DISC-FinLLM-13B exhibit subpar performance, potentially due to generating concise explanations not aligning with predicted answers. A close examination of the generated explanations indicates that, in many instances, these explanations do not align with the predicted answers. Through fine-tuning, all models show an improvement in generating more coherent explanations, achieving ∼20 BLEU-4 score and ∼40 ROUGE-L score.

4.3 Results of Application Assessment

Overall Performance. Table 5 provides an overview of the performance across the five application assessment tasks. GPT-4 emerges as the top performer, followed closely by GPT-4-turbo and ChatGPT. The performance gap among these three OpenAI LLMs is minimal. Among the open-source LLMs, Qwen-72B leads, followed by Qwen-14B, Qwen-7B, Baichuan2-13B, ChatGLM3-6B, Baichuan2-7B, Vicuna v1.5-7B, LlaMA2-70B, and LlaMA2-7B. Similar to the knowledge assessment results, the three financial LLMs lag behind their

Domain	Model	TC	MT				RE	RC	TG	Avg.
			En→Zh		Zh→En					
		Acc.	BLEU	COMET	BLEU	COMET	F1	R-L	R-L	
General	GPT-4-turbo	60.36 ± 0.10	**22.81 ± 0.08**	**79.89 ± 0.12**	19.90 ± 0.04	87.16 ± 0.20	**53.81 ± 0.29**	44.34 ± 0.13	24.22 ± 0.09	49.06
	GPT-4	**61.23 ± 0.03**	21.92 ± 0.03	78.32 ± 0.09	21.05 ± 0.02	87.20 ± 0.13	53.45 ± 0.09	46.34 ± 0.06	27.55 ± 0.05	**49.63**
	ChatGPT	52.42 ± 0.16	21.20 ± 0.12	78.21 ± 0.11	19.65 ± 0.08	86.82 ± 0.11	52.30 ± 0.19	**47.43 ± 0.11**	26.76 ± 0.06	48.10
	LLaMA2-7B	4.01 ± 0.04	1.59 ± 0.05	28.34 ± 0.14	3.37 ± 0.06	34.68 ± 0.18	21.48 ± 0.25	4.19 ± 0.03	1.09 ± 0.01	12.34
	LLaMA2-70B	16.67 ± 0.50	3.05 ± 0.06	43.19 ± 0.35	4.86 ± 0.02	40.59 ± 0.16	26.94 ± 0.28	7.07 ± 0.10	6.14 ± 0.15	18.56
	Vicuna v1.5-7B	22.77 ± 0.22	12.68 ± 0.14	56.39 ± 0.16	15.76 ± 0.07	79.51 ± 0.09	31.62 ± 0.05	42.56 ± 0.09	22.64 ± 0.01	35.49
	ChatGLM3-6B	27.65 ± 0.01	14.94 ± 0.07	62.40 ± 0.14	16.30 ± 0.63	78.26 ± 0.16	23.33 ± 0.20	43.08 ± 0.10	26.52 ± 0.13	36.56
	Baichuan2-7B	18.91 ± 0.25	18.78 ± 0.53	50.85 ± 0.11	18.11 ± 0.11	52.20 ± 0.07	23.29 ± 0.11	24.86 ± 0.04	15.46 ± 0.12	32.49
	Baichuan2-13B	15.06 ± 0.10	19.86 ± 0.07	74.44 ± 0.06	19.11 ± 0.11	84.15 ± 0.05	31.77 ± 0.10	43.45 ± 0.11	28.65 ± 0.00	39.56
	Qwen-7B	26.07 ± 0.62	18.10 ± 0.08	72.53 ± 0.13	19.27 ± 0.04	82.69 ± 0.11	35.15 ± 0.38	44.36 ± 0.05	28.00 ± 0.09	40.77
	Qwen-14B	39.87 ± 0.26	19.80 ± 0.11	74.99 ± 0.09	22.56 ± 0.06	84.81 ± 0.11	36.15 ± 0.12	45.20 ± 0.09	30.11 ± 0.08	44.18
	Qwen-72B	51.06 ± 0.20	22.08 ± 0.07	79.20 ± 0.03	**23.89 ± 0.03**	**87.21 ± 0.06**	49.21 ± 0.11	43.33 ± 0.05	**30.52 ± 0.02**	48.31
Financial	FinGPT-6B	19.10 ± 0.03	13.90 ± 0.12	60.64 ± 0.21	13.63 ± 0.08	73.48 ± 0.26	19.16 ± 0.24	39.75 ± 0.12	17.33 ± 0.05	32.12
	DISC-FinLLM-13B	23.24 ± 0.06	15.50 ± 0.13	70.95 ± 0.12	4.46 ± 0.05	80.63 ± 0.14	32.11 ± 0.29	43.32 ± 0.08	24.16 ± 0.10	36.80
	Tongyi-Finance-14B	29.91 ± 0.04	18.98 ± 7.63	73.84 ± 0.07	22.41 ± 1.87	84.61 ± 0.07	33.32 ± 0.16	45.00 ± 0.04	28.85 ± 0.02	42.12

Table 5: Comparison on the five application assessment tasks. TC, MT, RE, RC, and TG denote text classification, machine translation, relation extraction, reading comprehension, and text generation, respectively. Scores in **bold**/underline highlight the top/second-best performances. When calculating the average score, we first compute the average MT score by averaging BLEU and COMET scores, then use it to determine the overall average score.

general LLMs with similar parameter sizes.

Results of Text Classification. LLMs display varied text classification performance, strongly linked to model size. For example, OpenAI LLMs outperform open-source models by a significant margin, with a performance gap exceeding 20% in accuracy. Additionally, a substantial improvement from 26.07% to 51.06% is achieved by merely increasing the parameter size from 7B to 72B for Qwen. The performance gap for 6/7B LLMs (except LLaMA2-7B) is comparatively small.

Results of Machine Translation. Firstly, Qwen-7B/14B/72B performance is comparable to OpenAI LLMs in machine translation. Secondly, the performance trend for BLEU and COMET is not consistently aligned. For instance, Qwen-14B and Tongyi-Finance-14B achieve the higher BLEU scores in Chinese→English translation, but their COMET scores are significantly lower than those of OpenAI LLMs. DISC-FinLLM-13B attains a low BLEU score but a reasonable COMET score due to the translation involving numerous source-side words.

Results of Relation Extraction. Similar to text classification, LLMs display varied performance, with OpenAI models significantly outperforming Qwen-72B by $\sim 4\%$ in F1. Unlike text classification, there is no observed improvement by simply increasing the parameter size, as ChatGPT slightly underperforms GPT-4/GPT-4-turbo, while Qwen-7B performs similarly to Qwen-14B.

Results of Reading Comprehension. With the exception of LLaMA2-7B/-70B and FinGPT-6B, the performance gap among other LLMs is minimal, ranging from 42.56 to 47.43 in ROUGE-L. In contrast to other tasks, ChatGPT slightly outperforms GPT-4/GPT-4-turbo, while Qwen-7B performs similarly to Qwen-14B/-72B. We report BERTScore (Zhang et al., 2020) in Table 10 of Appendix B.

Results of Text Generation. OpenAI LLMs do not exhibit superiority over open-source LLMs, especially Qwen-14B, which achieves the best performance. Similar to reading comprehension, the performance gap among all LLMs, except LLaMA2-7B/-70B, Baichuan2-7B, and FinGPT-6B, is relatively small (22.64 to 30.52). Similarly, we report BERTScore in Table 10 of Appendix B.

For more detailed results per subtask in text classification, relation extraction, and text generation, please refer to Table 11, 12, and 13 in Appendex B.

5 Conclusion

In summary, this paper have presented CFLUE, comprehensive Chinese datasets in the financial domain designed for evaluating the performance of LLMs across diverse NLP tasks. With over 38K multiple-choice questions and 16K instances across generation tasks, CFLUE offers a robust benchmark for assessing LLM capabilities. The detailed results presented for several LLMs on CFLUE provide valuable insights into their performance. We hope that CFLUE will facilitate future advancements in LLMs, particularly in enhancing their proficiency of Chinese NLP tasks within the financial domain.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive feedback.

Ethics

CFLUE is a composite dataset sourced from various origins. Questions and corresponding solution explanations in the knowledge assessment section are derived from online mock exams. Instances in the application assessment are a combination of data from existing shared task datasets and information from select financial firms, annotated by our professional annotators.[20] We extend full credit to the original authors of each dataset incorporated into our CFLUE benchmark, and we have secured permissions for the inclusion of each dataset in CFLUE. The data obtained from our financial firm partners is also covered by our copyright. It is important to note that all datasets within CFLUE carry low ethical risks, with stringent measures in place to ensure the absence of any sensitive or personally identifiable information.

Limitations

Our work has several limitations. Firstly, we primarily use BLEU, COMET, ROUGE and BERTScore metrics to evaluate the performance of solution reasoning, machine translation, reading comprehension, and text generation. While informative, these metrics may not provide a comprehensive assessment of LLM outputs. Secondly, we employ identical or similar prompts for each group of subtasks in the application assessment, potentially hindering performance compared to prompts specifically tailored for each subtask. Lastly, our focus on the zero-shot setting when preparing prompts may limit LLMs from generating outputs as effectively as in a few-shot setting.

References

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 84–90.

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. *Computing Research Repository*, arXiv:2309.16609.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *Computing Research Repository*, arXiv:2309.10305.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, pages 38–45.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huan, and Zhongyu Wei. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *Computing Research Repository*, arXiv:2310.15205.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledg, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*, pages 3697–3711.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL*, pages 2368–2378.

Duxiaoman-DI. 2022. Financeiq. https://huggingface.co/datasets/Duxiaoman-DI/FinanceIQ.

FinSBD3. 2021. Financial sdb 3. https://sites.google.com/nlg.csie.ntu.edu.tw/finweb2021/shared-task-finsbd-3.

FiQA. 2018. Financial question answering. https://sites.google.com/view/fiqa.

[20] Prior to annotation, we performed anonymization procedures on non-public financial data obtained from financial firms.

Cuiyun Han, Jinchuan Zhang, Xinyu Li, Guojin Xu, Weihua Peng, and Zengfeng Zeng. 2022. Duee-fin: A large-scale dataset for document-level event extraction. In *Proceedings of NLPCC*, pages 172–183.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

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. In *Proceedings of NeurIPS*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Computing Research Repository*, arXiv:2311.11944.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *Computing Research Repository*, arXiv:2311.06602.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *Computing Research Repository*, arXiv:2306.09212.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and Eric Zelikman. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Chin-Yew Lin and Eduard Hovy. 2002. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of ACL*, page 311318.

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. CSDS: A fine-grained Chinese dataset for customer service dialogue summarization. In *Proceedings of EMNLP*, pages 4436–4451.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, pages 3214–3252.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset. In *Proceedings of NeurIPS*.

Macedo Maia, Siegfried Handschuh, Andr é Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of WWW*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the American Society for Information Science and Technology*, 65:782–796.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*, page 311318.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don ' t know: Unanswerable questions for squad. In *Companion Proceedings of WWW*, pages 784–789.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, pages 2685–2702.

Junxiang Ren, Sibo Wang, Ruilin Song, Yuejiao Wu, Yizhou Gao, Borong An, Zhen Cheng, and Guoqiang Xu. 2022. Iree: A fine-grained dataset for chinese event extraction in investment research. In *Proceedings of China Conference on Knowledge Graph and Semantic Computing*, pages 205–210.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64:99106.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of EMNLP*, pages 2322–2335.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results. *Computing Research Repository*, arXiv:2009.04202.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adri à Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tianchi. 2019. Ccks2019 financial domain document-level event entity extraction dataset. https://tianchi.aliyun.com/dataset/111237.

Tianchi. 2020. Ccks2020 financial domain document-level event entity extraction dataset. https://https://tianchi.aliyun.com/dataset/111209.

Tianchi. 2021. Ccks2021 financial domain event causal relationship extraction dataset. https://tianchi.aliyun.com/dataset/110901.

Tianchi. 2022. Ccks2022 financial domain few-shot event extraction dataset. https://tianchi.aliyun.com/dataset/136800.

TongyiFinance. 2023. Tongyi-finance-14b-chat. https://www.modelscope.cn/models/TongyiFinance/Tongyi-Finance-14B-Chat/summary.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth é e Lacroix, Baptiste Rozi è re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Computing Research Repository*, arXiv:2302.13971.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of NeurIPS*, pages 3266–3280.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of COLING*, pages 4762–4772.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of ACL*, pages 4791–4800.

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. *Computing Research Repository*, 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. *Computing Research Repository*, arXiv:2308.09975.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

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. *Computing Research Repository*, arXiv:2304.06364.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of ACL-IJCNLP*, pages 3277–3287.

A Detailed Statistics of CFLUE

Table 6 lists the subjects in the knowledge assessment, as well as the number of questions included in each subject. Figure 3 shows the subtasks in the application assessment, as well as the number of instances included in each subtask.

B Detailed Performance of CFLUE

For knowledge assessment, Table 7 and Table 8 show the performance of answer prediction in accuracy per subject and per question type, respectively. Figure 4 shows the performance curves concerning input length. It shows that LLMs generally show higher accuracy with input lengths between 60 and 90. Interesting, their performance varies when the input is shorter with less then 30 tokens.

{

ID: 2327

名称: 中级经济师/Intermediate Economics Professional Qualification

task: 单项选择题/multiple-choice question with a single correct option

question: 下列选项中，对著作权登记的效力描述正确的是（）。/ Among the following options, the correct description of the effectiveness of copyright registration is ().

choices: {A: 著作权产生的要件/ The prerequisites for the creation of copyright,

B: 著作权转让的生效要件/ The conditions for the effectiveness of copyright transfer,

C: 初步证据/ Preliminary evidence,

D: 著作权许可的要件/ The requirements for copyright licensing}

answer: C

analysis: 作品不论是否登记，作者或者其他著作权人依法取得的著作权均不受影响。但著作权登记有助于解决因著作权归属造成的著作权纠纷，并为解决著作权纠纷提供初步证据。这一初步证据可以因相反证据而被推翻。/ The rights obtained by authors or other copyright holders in accordance with the law are not affected, whether or not the work is registered. However, copyright registration can help resolve disputes arising from the ownership of copyright and provides preliminary evidence for resolving copyright disputes. This preliminary evidence can be overturned by contradictory evidence.

}

Figure 2: An example of our annotated multi-choice question. English translations are provided along the corresponding Chinese text.

It is important to highlight that there has been extensive research on quantitative reasoning, particularly in the context of program synthesis (Koncel-Kedziorski et al., 2023; Islam et al., 2023). Our set of multiple-choice questions specifically contains quantitative reasoning questions. We utilize heuristic rules, followed by human validation, to identify quantitative reasoning questions within our test set. Consequently, we identify 334 quantitative reasoning questions, and their performance is detailed in Table 9. The results indicate that the performance on quantitative reasoning questions is significantly lower than the overall performance, suggesting the challenges that LLMs face in answering such questions.

For application assessment, Table 11, 12, and 13 show the performance per subtasks in text classification, relation extraction, and text generation, respectively.

Figure 5 compares the performance of models in application assessment tasks.

C Prompt Examples

Figure 6 to 11 illustrate the evaluation prompts used in our experimentation.

Subject	Size			
	Train	Valid	Test	All
基金从业资格 / Asset Management Association of China	2,691	336	337	3,364
金融理财师 / Associate Financial Planner	1,166	146	146	1,458
会计从业资格 / Certificate of Accounting Professional	2,273	284	283	2,840
银从中级资格 / Certification of China Banking Professional (Intermediate)	3,395	424	424	4,243
银行初级资格 / Certification of China Banking Professional (Preliminary)	3,617	453	452	4,522
期货从业资格 / Certificate of Futures Qualification	1,862	233	232	2,327
证券从业资格 / Certification of Securities Professional	1,223	153	154	1,530
中国精算师 / Certified China Actuary	103	13	13	129
注册会计师 / Certified Public Accountant	3,644	456	456	4,556
保险从业资格 / China Insurance Certification & Education	557	69	70	696
反假货币考试 / Counterfeit Currency Detection Exam	486	60	61	607
黄金从业资格 / Gold Trading Qualification Certificate	679	85	85	849
中级经济师 / Intermediate Economics Professional Qualification	4,968	622	620	6,210
初级经济师 / Junior Economics Professional Qualification	3,279	409	411	4,099
证券专项考试 / Securities Special Examination	965	121	120	1,206
Total	30,908	3,864	3,864	38,636

Table 6: Detailed statistics of the knowledge assessment of CFLUE.

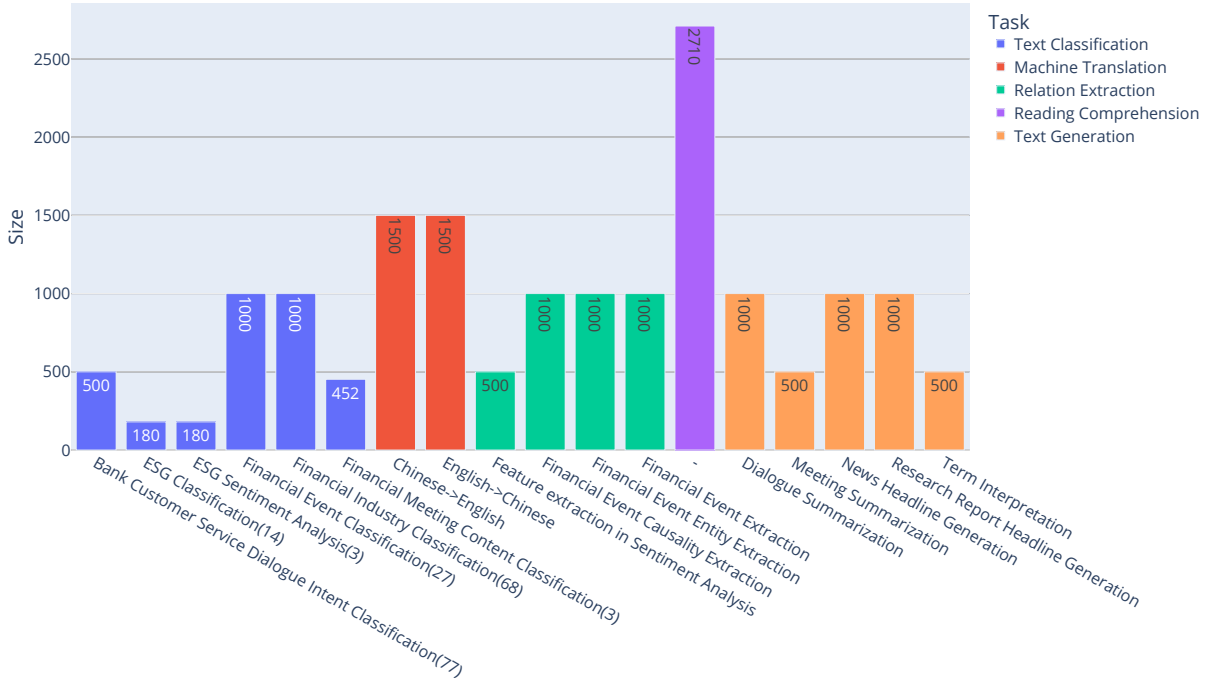

Figure 3: Detailed statistics of the application assessment of CFLUE. For subtasks in text classification, the numbers within brackets represent the respective number of classes.

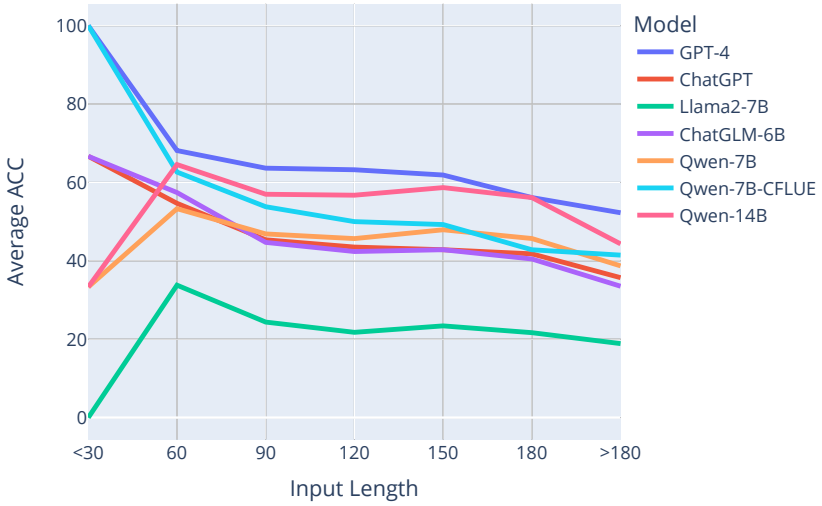

Figure 4: Performance in accuracy over different input lengths in knowledge assessment.

Subject	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Qwen-7b-CFLUE	Random
Asset Management Association of China	60.87	71.51	45.7	55.38	49.84	26.40
Associate Financial Planner	63.01	34.25	41.10	51.37	51.37	23.97
Certificate of Accounting Professional	57.45	36.88	42.20	51.06	54.06	24.32
Certification of China Banking Professional (Intermediate)	61.08	45.75	49.12	51.65	51.29	21.65
Certification of China Banking Professional (Preliminary)	67.04	52.21	49.12	61.95	50.66	25.44
Certificate of Futures Qualification	57.94	45.92	42.49	48.93	52.36	25.49
Certification of Securities Professional	53.25	44.81	49.35	54.55	54.25	22.59
Certified China Actuary	30.77	53.87	38.46	38.46	53.85	23.07
Certified Public Accountant	49.78	34.43	38.60	48.68	48.02	20.96
China Insurance Certification	64.29	51.43	57.14	62.86	46.38	24.0
Counterfeit Currency Detection Exam	57.38	40.98	42.63	57.38	55.74	28.52
Gold Trading Qualification Certificate	50.59	45.88	41.18	45.88	49.41	28.84
Intermediate Economics Professional Qualification	61.67	40.26	43.16	52.66	51.04	15.84
Junior Economics Professional Qualification	62.93	43.90	47.07	61.71	53.66	17.56
Securities Special Examination	69.17	40.83	47.50	65.00	51.24	21.16

Table 7: Performance of answer prediction in accuracy per subject. To save space, we list the performance of five LLMs.

Question Type	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Qwen-7b-CFLUE	Random
单项选择题 / with single answer	67.70	49.36	54.01	64.89	51.98	25.19
多项选择题 / with multiple answers	44.83	27.58	27.07	34.50	50.26	6.79
判断题 / of true or false	68.77	55.87	52.44	61.60	55.01	50.60

Table 8: Performance of answer prediction in accuracy per question type. To save space, we list the performance of five LLMs.

Type	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Qwen-7b-CFLUE
quantitative reasoning	46.41	26.95	32.63	32.04	34.13

Table 9: Performance of answer prediction in accuracy for quantitative reasoning questions. To save space, we list the performance of five LLMs.

Domain	Model	Reading Comprehension	Text Generation
General	GPT-4-turbo	74.20	69.28
	GPT-4	75.49	70.01
	ChatGPT	**75.82**	69.26
	LLaMA2-7B	47.84	45.96
	LLaMA2-70B	50.19	50.11
	Vicuna v1.5-7B	68.98	63.88
	ChatGLM3-6B	73.54	68.55
	Baichuan2-7B	62.37	58.29
	Baichuan2-13B	73.22	70.02
	Qwen-7B	74.25	69.62
	Qwen-14B	74.79	70.84
	Qwen-72B	73.92	**71.22**
Financial	FinGPT-6B	71.30	48.71
	DISC-FinLLM-13B	73.86	66.66
	Tongyi-Finance-14B	74.57	70.06

Table 10: Performance in BERTScore for reading comprehension and text generation in application assessment.

SubTask	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Baichuan2-7B
Bank Customer Service Dialogue Intent Classification	61.85	56.00	31.00	43.00	0.40
ESG Classification	44.44	42.22	12.78	20.45	48.33
ESG Sentiment Analysis	70.11	41.11	68.33	62.50	74.44
Financial Event Classification	71.29	52.00	32.30	42.40	18.70
Financial Industry Classification	40.77	41.90	13.10	21.60	0.40
Financial Meeting Content Classification	82.22	74.78	15.49	54.55	41.59

Table 11: Performance in accuracy per subtask in text classification. To save space, we list the performance of five LLMs.

SubTask	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Baichuan2-7B
Feature extraction in Sentiment Analysis	45.15	36.42	14.20	27.59	17.35
Financial Event Causality Extraction	27.77	27.26	13.39	20.55	10.10
Financial Event Entity Extraction	88.09	89.20	87.60	85.29	74.90
Financial Event Extraction	53.62	48.36	15.93	5.63	17.01

Table 12: Performance in F1 score per subtask in relation extraction. To save space, we list the performance of five LLMs.

SubTask	GPT4	ChatGPT	Qwen-7B	Qwen-14B	Baichuan2-7B
Dialogue Summarization	37.68	37.04	36.25	38.06	36.72
Meeting Summarization	35.29	33.04	33.57	34.75	31.52
News Headline Generation	21.95	21.45	21.66	22.33	23.81
Research Report Headline Generation	23.78	23.62	23.32	24.64	24.77
Term Interpretation	18.30	16.01	18.87	18.55	13.89

Table 13: Performance in ROUGE-L per subtask in text generation. To save space, we list the performance of five LLMs.

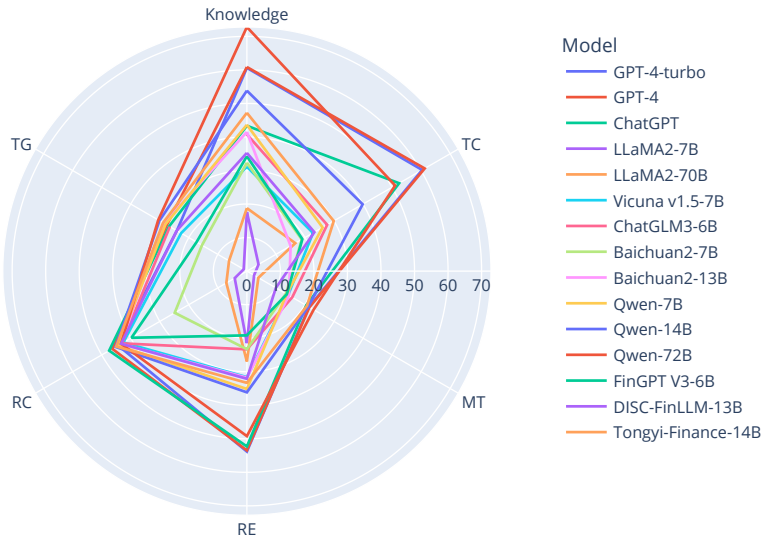

Figure 5: Comparison among LLMs in subtasks of knowledge and application assessment.

假设你是一位金融行业专家，请回答下列问题。注意：题目是单选题，只需要返回一个最合适的选项，若有多个合适的答案，只返回最准确的即可。注意：结果只输出两行，第一行只需要返回答案的英文选项(注意只需要返回一个最合适的答案)，第二行进行简要的解析，输出格式限制为："答案："，"解析："。
Assuming you are a financial industry expert, please answer the following questions. Note: The question is a multiple-choice question, and only the most appropriate option needs to be returned. If there are multiple appropriate answers, only the most accurate answer will be returned. Note: The result only outputs two lines. The first line only needs to return the English option of the answer (note that only the most appropriate answer needs to be returned). The second line performs a brief analysis. The output format is limited to: "Answer:", "Explanation:".

甲公司处于可持续增长状态。预计20×1年末净经营资产1000万元，净负债340万元，20×1年营业收入1100万元，净利润100万元，股利支付率40%。甲公司20×1年可持续增长率是（ ）。
Company A is in a state of sustainable growth. It is estimated that the net operating assets at the end of 20×1 will be 10 million yuan, the net liabilities will be 3.4 million yuan, the operating income in 20×1 will be 11 million yuan, the net profit will be 1 million yuan, and the dividend payout rate will be 40%. Company A's sustainable growth rate in 20×1 is ( ).

A. 6.5%
B. 10%
C. 9.1%
D. 6.1%

答案：B
Answer: B
解析：可持续增长率＝本期利润留存/(期末股东权益-本期利润留存)×100%期末股东权益＝期末净经营资产－期末净负债＝1000-340＝660（万元）利润留存＝100×（1-40%）＝60（万元）所以，可持续增长率＝60/（660－60）×100%＝10%
Explanation: Sustainable growth rate = Retained earnings / (Ending shareholders' equity - Retained earnings) × 100%. Ending shareholders' equity = Ending net operating assets - Ending net debt = 1000 - 340 = 660 (thousand yuan). Retained earnings = 100 × (1 - 40%) = 60 (thousand yuan). Therefore, the sustainable growth rate = 60 / (660 - 60) × 100% = 10%.

Figure 6: Illustration of answer prediction and reasoning. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.

你是一个ESG专家，阅读给定的ESG报告的内容，将其分成以下几个类别：公共政策,经济绩效,物料,能源,当地劳工及社群权益,非直接经济影响,反不正当竞争,顾客健康与安全,依法合规纳税,市场占有率,安全管理实践,采购行为,职业培训及教育,排放.
You are an ESG (Environmental, Social, and Governance) expert. Please read the given ESG report and categorize its contents into the following categories: Public Policy, Economic Performance, Materials, Energy, Local Labor and Community Rights, Indirect Economic Impact, Anti-Unfair Competition, Customer Health and Safety, Legal and Tax Compliance, Market Share, Safety Management Practices, Procurement Practices, Occupational Training and Education, Emissions.

ESG报告内容:
ESG Report Content:

(省略)
(Content omitted)

所属类别：公共政策
Category: Public Policy

Figure 7: Illustration of ESG classification subtask in text classification. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.

你是一个金融行业专家，请将下面的英文翻译成准确并专业的中文。
You are a financial industry expert. Please translate the following English into accurate and professional Chinese.

英文:
English:
If Jamie Dimon fulfils the plan he announced this week, and remains chairman and chief executive of JPMorgan Chase until 2023, he will have been at the helm for 17 years.

中文:
Chinese:
如果杰米·戴蒙(Jamie Dimon)执行他最近宣布的计划，继续担任摩根大通(JPMorgan Chase)董事长兼首席执行官至2023年，那他将执掌摩根大通帅印17年。

Figure 8: Illustration of English→Chinese subtask in machine translation. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.

你是一个金融行业专家，请帮我完成一个抽取+情感分析的复合任务。请阅读下面股评内容，从中抽取行业及其对应的情感分类(正向、负向、中性)，注意股评中可能包含了多个行业的看法，不要遗漏。
You are a financial industry expert; please help me complete a composite task of extraction and sentiment analysis. Please read the following stock review content and extract the industry along with its corresponding sentiment classification (positive, negative, neutral). Note that the stock review may contain opinions on multiple industries, so please do not miss any.

输出样例：
Sample Output:
汽车: 正向
Automobiles: Positive
白酒: 负向
Liquor: Negative

股评内容: (内容略)
Stock Review Content: (content omitted)

抽取结果：
Extraction Result:
新能源： 正向
New Energy: Positive

Figure 9: Illustration of ESG sentiment analysis subtask in relation extraction. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.

你是一专业的金融分析师，仔细阅读并理解下面的研报内容，并回答问题：
You are a professional financial analyst. Please read and thoroughly understand the following research report content and answer the questions:

研报内容：
The content of the research report:
(内容略)
(Content omitted)

基于对上面内容的理解回答问题，你的回答尽可能简洁。严格根据研报内容进行回答，不要编造事实。如果根据研报内容无法回答相关问题，直接回答“我不知道”。现在开始：
Based on your understanding of the above content, answer the questions as concisely as possible. Answer strictly based on the content of the research report, avoiding fabrication. If you cannot answer a question based on the content of the research report, respond with "I don't know." Let's begin now:

问题: (问题略)
Question: (question omitted)

答案: (答案略)
Answer: (answer omitted)

Figure 10: Illustration of reading comprehension task. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.

你是一个客服对话摘要生成助手，请根据客服的对话内容，生成一段客服对话摘要，尽量精炼可以概括主要内容。
You are a customer service dialogue summary generation assistant. Please generate a concise summary of the customer service conversation based on its content, focusing on summarizing the main points

对话内容:
Conversation :
(内容略)
(Content omitted)

对话摘要:
Dialogue summarization:
(内容略)
(Content omitted)

Figure 11: Illustration of dialogue summarization subtask in text generation. The model generates the red text as an autocompleted response, while the preceding text serves as the input prompt. English translations are provided below the corresponding Chinese text.