CFinBench: A Comprehensive Chinese Financial Benchmark for Large Language Models

Ying Nie[1*] **Binwei Yan**[1*] **Tianyu Guo**[1] **Hao Liu**[1] **Haoyu Wang**[1] **Wei He**[1]
Binfan Zheng[2] **Weihao Wang**[3] **Qiang Li**[3] **Weijian Sun**[2] **Yunhe Wang**[1†] **Dacheng Tao**[4]
[1]Huawei Noah's Ark Lab [2]Huawei GTS [3]Huawei Group Finance
[4]Nanyang Technological University
{ying.nie, yanbinwei, tianyu.guo, yunhe.wang}@huawei.com

Abstract

Large language models (LLMs) have achieved remarkable performance on various NLP tasks, yet their potential in more challenging task like finance, has not been fully explored. In this paper, we present CFinBench: a meticulously crafted, the most comprehensive evaluation benchmark to date, for assessing the financial knowledge of LLMs under Chinese context. In practice, to better align with the career trajectory of Chinese financial practitioners, we build a systematic evaluation from 4 first-level categories: (1) *Financial Subject*: whether LLMs can memorize the necessary basic knowledge of financial subjects, such as economics, statistics and auditing. (2) *Financial Qualification*: whether LLMs can obtain the needed financial qualified certifications, such as certified public accountant, securities qualification and banking qualification. (3) *Financial Practice*: whether LLMs can fulfill the practical financial jobs, such as tax consultant, junior accountant and securities analyst. (4) *Financial Law*: whether LLMs can meet the requirement of financial laws and regulations, such as tax law, insurance law and economic law. CFinBench comprises 99,100 questions spanning 43 second-level categories with 3 question types: single-choice, multiple-choice and judgment. We conduct extensive experiments on a wide spectrum of representative LLMs with various model size on CFinBench. The results show that GPT4 and some Chinese-oriented models lead the benchmark, with the highest average accuracy being 66.02%, highlighting the challenge presented by CFinBench. All the data and evaluation code are open sourced at https://cfinbench.github.io/.

1 Introduction

Recently, there has been a significant advancement in LLMs, exemplified by the representative models like ChatGPT (OpenAI, 2021), GPT4 (OpenAI, 2023), LLaMA (Touvron et al., 2023a,b; Meta, 2024), Baichuan (Yang et al., 2023a), InternLM (Team, 2023) and ChatGLM (Zeng et al., 2022), *etc*. At the same time, the corresponding evaluation works for LLMs are flourishing and a series of evaluation benchmarks have been proposed like MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2024), Xiezhi (Gu et al., 2024) and AGIEval (Zhong et al., 2023), *etc*. These benchmarks have been instrumental in catalyzing the progress of LLMs, as they enable the quantitative assessment of advanced knowledge and complex reasoning abilities.

Finance is the backbone of modern society, playing a vital role in facilitating economic growth and prosperity (Shiller, 2013). However, mastering the intricacies of financial knowledge is challenging for individuals, due to its intricate nature and dynamic environment. Therefore, endowing LLMs with financial knowledge is essential as it can provide significant convenience and insight to humanity. For example, BloombergGPT (Wu et al., 2023), which possesses 50 billion parameters, exhibits superior performance across multiple financial tasks. Similarly, FinMA (Xie et al., 2023) is crafted by fine-tuning LLaMA (Touvron et al., 2023a) with the financial instruction data. Of course, a comprehensive financial evaluation benchmark is also essential for financial LLMs.

Several benchmarks have been introduced for better evaluating financial LLMs. FLUE (Shah et al., 2022) first introduces the financial benchmark across 5 NLP tasks in English domain, and its successor, FLARE (Xie et al., 2023), further extends it with financial time-series reasoning task like stock price movements forecasting. In addition to English domain, benchmarks in Chinese are another one of significant importance. BBT-CFLEB (Lu et al., 2023) presents the first Chinese financial evaluation benchmark, which includes 6

[*]Equal Contribution
[†]Corresponding Author

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 876–891

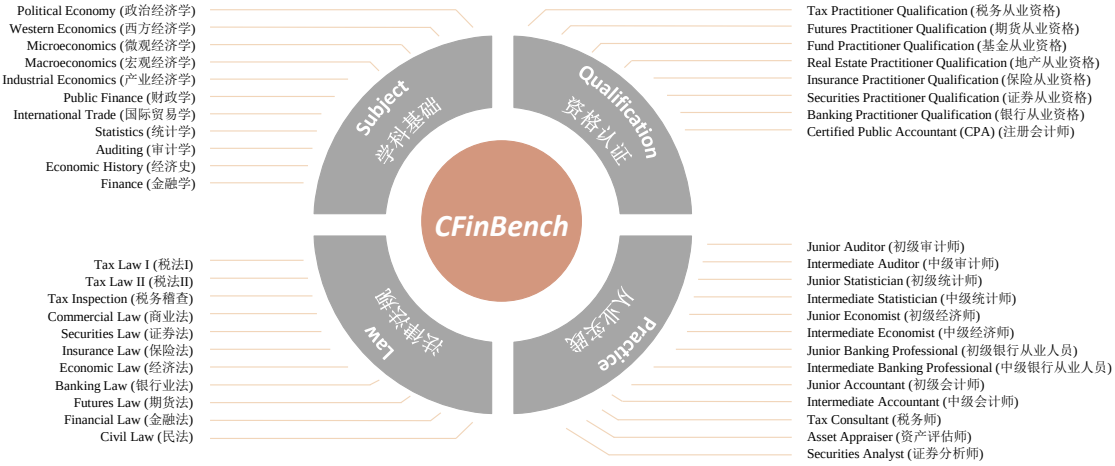


Figure 1: CFinBench comprises 4 first-level categories and 43 second-level categories, which are more align with the career trajectory of Chinese financial practitioners.

datasets covering both understanding and generation tasks. FinEval (Zhang et al., 2023a) builds a collection of 4,661 single-choice questions including 4 first-level categories: finance, economy, accounting, and certificate. However, FinEval is constrained by its limited size, and the category coverage is also inadequate for capturing the real-world financial scenarios. Moreover, the BBT-CFLEB, which targets the basic NLP tasks in finance, struggles to provide sufficient challenges for the increasingly advanced large language models.

In this paper, we present CFinBench: a meticulously crafted, the most comprehensive evaluation benchmark to date, for assessing the capabilities of LLMs on Chinese financial tasks. The design philosophy of our benchmark aligns with the career progression trajectory of financial practitioners, which can be likened to a 'leveling up' in a game. Specifically, it begins with mastering the required foundational knowledge in financial subjects, followed by obtaining the necessary qualified certifications, and subsequently honing skills through practical experience in industry applications. Last but not least, compliance with financial laws and regulations is also a crucial aspect. As described in Figure 1, we include 4 first-level categories: (1) *Financial Subject*: examining whether LLMs can memorize the foundational knowledge in financial subjects, such as political economy, statistics and macroeconomics, *etc*. (2) *Financial Qualification*: examining whether LLMs can obtain the necessary qualified certifications for financial practitioners, such as tax practitioner qualification, futures practitioner qualification and fund practitioner qualification, *etc*. (3) *Financial Practice*: examining whether LLMs can fulfill the specific tasks in financial jobs, such as junior banking professional, asset appraiser and junior statistician, *etc*. (4) *Financial Law*: examining whether LLMs can comply with the financial laws and regulations, such as securities law, insurance law and economic law, *etc*. CFinBench are primarily sourced from the mock exams and financial reports freely available on the Internet. To enhance the quality and diversity of the benchmark, and mitigate the problem of data contamination, we perform a series of rigorous data processing pipelines, encompassing data cleaning, internal and external de-duplication, LLM-assisted rephrasing, option shuffling, and multi-round human-in-the-loop cross-validation. CFinBench comprises 99,100 questions spanning 43 second-level categories with 3 question types: single-choice, multiple-choice and judgment.

We conduct extensive experiments on a wide spectrum of representative LLMs with various model size on CFinBench. The results show that GPT4 and some Chinese-oriented models like Qwen (Bai et al., 2023), Yi (Young et al., 2024), and XuanYuan (Zhang and Yang, 2023), *etc*. lead the benchmark, with the highest average accuracy being 66.02%, highlighting the challenge presented by CFinBench. It also indicates that there is still significant room for improvement in current Large Language Models (LLMs) within the Chinese financial domain.

2 Related Work

Financial LLMs The advent of ChatGPT (OpenAI, 2021) marks a significant milestone in natural language processing (NLP), demonstrating

Table 1: Comparison of the proposed CFinBench with other Chinese-oriented financial benchmarks.

Benchmark	#Test Questions	#Categories	#Question Types	Task
BBT-CFLEB (Lu et al., 2023)	20,416	6	4	Basic NLP
CGCE (Zhang et al., 2023b)	150	4	1	QA
CFBenchmark (Lei et al., 2023)	3,917	8	3	Basic NLP
FinEval (Zhang et al., 2023a)	4,661	34	1	Advanced Knowledge
FinanceIQ (Zhang and Yang, 2023)	7,173	36	1	Advanced Knowledge
Ant-Fin-Eva (Group, 2023)	8,445	33	1	Advanced Knowledge
CFLUE (Zhu et al., 2024)	16,522	33	6	Advanced Knowledge & Application
CFinBench	99,100	43	3	Advanced Knowledge

the remarkable capabilities of large language models with billions of parameters across a diverse range of tasks. This progress is further amplified by the release of GPT4 (OpenAI, 2023), which exhibits even greater generalization abilities. There have also been studies that are dedicated to adapting LLMs to the financial domain. BloombergGPT (Wu et al., 2023), the first proprietary LLM comprising 50 billion parameters, has been meticulously tailored for the financial domain. The successors like FinGPT (Yang et al., 2023b) and FinMA (Xie et al., 2023) introduce the financial LLMs based on fine-tuning LLaMA, by means of low-rank adaptation or full-parameters. Also, the Chinese-oriented financial LLMs like DISC-FinLLM (Chen et al., 2023a), CFGPT (Li et al., 2023b), XuanYuan (Zhang and Yang, 2023) and YunShan (Wang et al., 2023), *etc*. have also demonstrated excellent performance across multiple Chinese financial tasks. At the same time, to objectively and quantitatively measure the capabilities of LLMs, a comprehensive and thorough evaluation benchmark is crucial.

Financial Benchmarks FLUE (Shah et al., 2022) first introduces a financial evaluation benchmark across 5 NLP tasks in English context, and its successor, FLARE (Xie et al., 2023), further includes financial time-series reasoning task like stock movement prediction. The FinBen (Xie et al., 2024) reorganizes 35 public English datasets across 23 financial tasks into three spectrums of difficulty. For the Chinese domain, BBT-CFLEB (Lu et al., 2023) presents the first Chinese financial evaluation benchmark, which includes 6 datasets covering both understanding and generation tasks. By integrating FLARE (Xie et al., 2023) and BBT-CFLEB (Lu et al., 2023), ICE-FLARE (Hu et al., 2024) enables the evaluation of bilingual financial tasks. CFBenchmark (Lei et al., 2023) assesses the text processing capabilities across recognition, classification, and generation tasks. CGCE (Zhang et al., 2023b) incorporates 200 general and 150 finance-specific question-answering questions. The recent study of CFLUE (Zhu et al., 2024) builds the datasets tailored for both knowledge assessment and application assessment. By the way, the most similar works to ours are FinanceIQ (Zhang and Yang, 2023) and FinEval (Zhang et al., 2023a). FinanceIQ (Zhang and Yang, 2023) encompasses 10 primary categories, including economists, securities professionals and actuaries, *etc*. with a total of 36 subcategories and 7,173 single-choice questions. FinEval (Zhang et al., 2023a) includes four primary categories, *i.e*., finance, economy, accounting, and certificate, with a total of 34 subcategories and 4,661 single-choice questions. However, the two benchmarks are limited in size. In contrast, the proposed CFinBench is more comprehensive, comprising a more reasonable dimension of financial abilities, with a total of 99,100 questions across 3 question types. The detailed comparison of the CFinBench with other Chinese-oriented financial benchmarks is summarized in Table 1. It is worth highlighting that we are not a aggregation of existing benchmarks, but rather a significant supplement from the perspective of professional practitioners in terms of advanced knowledge and complex reasoning, which, in conjunction with preceding benchmarks (Shah et al., 2022; Lu et al., 2023; Zhang et al., 2023a; Lei et al., 2023; Zhu et al., 2024), foster the development of financial LLMs.

3 CFinBench

In this section, we first introduce the overall design principle of the proposed benchmark, then we elaborate the taxonomy of CFinBench including financial subject, financial qualification, financial practice, and financial law. The detailed process of data construction including data sources, data processing and human cross-validation is also presented at the end.

在下列关于股利政策的表述中，股利与利润之间保持固定比例关系，体现风险投资与风险收益对等关系的是：
In the following statements about dividend policy, a fixed proportional relationship is maintained between dividends and profits, reflecting the relationship between risky investments and risk-return parity:
A. 固定股利支付率政策
Fixed dividend payout ratio policy
B. 剩余股利政策
Residual dividend policy
C. 低正常股利加额外股利政策
Low normal dividend plus bonus dividend policy
D. 固定股利政策
Fixed dividend policy
答案：A
Answer: A

(a) Example of single-choice questions in Industrial Economics.

某商业银行某日营业终了，业务库存现金300万元，在央行的存款准备金账户余额1800万元，其中法定存款准备金800万元。该商业银行营运过程中，因同业往来引起头寸减少的因素为 ()。
At the end of business on a certain date, a commercial bank has 3 million yuan in cash on hand for operations and a reserve account balance of 18 million yuan at the central bank, of which 8 million yuan is legal reserve. In the course of the commercial bank's operations, the factors that caused the position to decrease due to interbank transactions is ( ).
A. 由他行调入资金 Transfer of funds from other banks
B. 同业往来利息支出 Interest expense on interbank transactions
C. 上级行调走资金 Funds transferred from parent bank
D. 偿还中央银行借款 Repayment of central bank loans
E. 贷款信用评估 Loan Credit Assessment
答案：B C
Answer: B C

(b) Example of multiple-choice questions in Junior Economist.

《中华人民共和国商业银行法》规定“商业银行贷款，借款人应当提供担保”。因此，即使借款人资信优良，有足够的偿还贷款能力，也必须提供担保。
The Law of the People's Republic of China on Commercial Banks stipulates that "the borrower shall provide security for loans made by commercial banks". Therefore, even if the borrower has excellent creditworthiness and sufficient ability to repay the loan, he or she must provide a guarantee.
答案：错误
Answer: Wrong

(c) Example of judgment questions in Banking Law.

Figure 2: Examples of 3 types of questions. English translations are shown in blue for better readability.

3.1 Overview

The motivation of CFinBench is to evaluate the financial knowledge of large language models in the context of Chinese. Inspired by predecessors (Huang et al., 2024; Fei et al., 2023; Zhang et al., 2023a), we also focus on the advanced knowledge and complex reasoning abilities, which, compared to traditional NLP capabilities, pose a greater challenge to the increasingly advanced LLMs of today. In practice, we build CFinBench based on the real-world examination questions used in China for assessing financial professionals across multiple dimensions. We include 3 question types: single-choice, multiple-choice and judgment, as exemplified in Figure 2. Compared with the single-choice questions alone in most existing works (Zhang et al., 2023a; Zhang and Yang, 2023; Clark et al., 2018; Group, 2023; Huang et al., 2024), a broader range of question types can more comprehensively assess the capabilities of LLMs. Specifically, for single-choice questions, each question has four options, with only one correct answer. For multiple-choice questions, each question has four or five options, with at least two correct answers. For judgment questions, each question requires a direct judgment of whether the statement is correct or wrong.

3.2 Taxonomy

In the evaluation of large language models, a diverse array of tasks is often preferred to comprehensively assess their capabilities. A hierarchical evaluation framework enables a more nuanced understanding of the abilities of LLMs. Instead of categorizing the financial tasks solely based on their subjects (Zhang et al., 2023a; Zhang and Yang, 2023), we thoroughly explore the characteristics of Chinese financial system, and reorganize the financial tasks into more reasonable categories. Specifically, the process starts with acquiring fundamental knowledge in financial subjects, followed by passing essential professional qualifications, and subsequently refining skills through practical experience in industry applications. Additionally, adherence to laws and regulations is also a critical aspect. In practice, we include 4 first-level categories and 43 second-level categories, which are summarized in Figure 1.

- **Financial Subject**: The purpose of the financial subject is to test whether LLMs can memorize the essential foundational knowledge in financial subjects. Specifically, 11 subjects are included: Political Economy, Western Economics, Microeconomics, Macroeconomics, Industrial Economics, Public Finance, International Trade, Statistics, Auditing, Economic History, and Finance. These subjects provide a comprehensive framework for understanding the intricacies of economic systems, market structures, and financial institutions.

- **Financial Qualification**: The objective of the financial qualification is to examine whether LLMs can obtain necessary qualified certifications for finance professionals. We include 8 qualifications: Tax Practitioner Qualification, Futures Practitioner Qualification, Fund Practitioner Qualification, Real Estate Practitioner

Qualification, Insurance Practitioner Qualification, Securities Practitioner Qualification, Banking Practitioner Qualification, and Certified Public Accountant (CPA). By obtaining these qualifications, professionals can enhance their knowledge and skills in areas such as financial analysis, risk management, and financial planning.

- **Financial Practice**: Financial practice is to examine whether LLMs can fulfill the specific tasks in practical financial jobs. We include 13 jobs: Junior/Intermediate Auditor, Junior/Intermediate Statistician, Junior/Intermediate Economist, Junior/Intermediate Banking Professional, Junior/Intermediate Accountant, Tax Consultant, Asset Appraiser, and Securities Analyst. These practices involve the application of financial concepts and techniques to real-world problems, requiring professionals to possess a deep understanding of financial markets, instruments, and regulations.

- **Financial Law**: The purpose of financial law is to test whether LLMs can comply with financial laws and regulations. Specifically, it includes 11 exams of laws: Tax Law I/II, Tax Inspection, Commercial Law, Securities Law, Insurance Law, Economic Law, Banking Law, Futures Law, Financial Law and Civil Law. These laws provide the legal foundation for financial transactions, investments, and operations. Proficiency in financial laws can reduce the occurrence of illegal activities.

3.3 Data Construction

3.3.1 Data Sources

Our dataset is primarily sourced from publicly accessible channels like mock exams on internet, public books on economics and law, announcements and financial reports of listed companies, financial articles and news, *etc*.

3.3.2 Data Processing

The collected data come in various formats, including PDF, EPUB, Microsoft Word documents and web pages. Documents in PDF format and EPUB format are parsed into text using PyMuPDF[1] and EbookLib[2] respectively. We standardized all single-choice questions to have exactly four options, excluding those with fewer options and randomly removing excess wrong options from those with more than four. Similarly, for multiple-choice questions, to maintain uniformity, we only retain questions with four or five options.

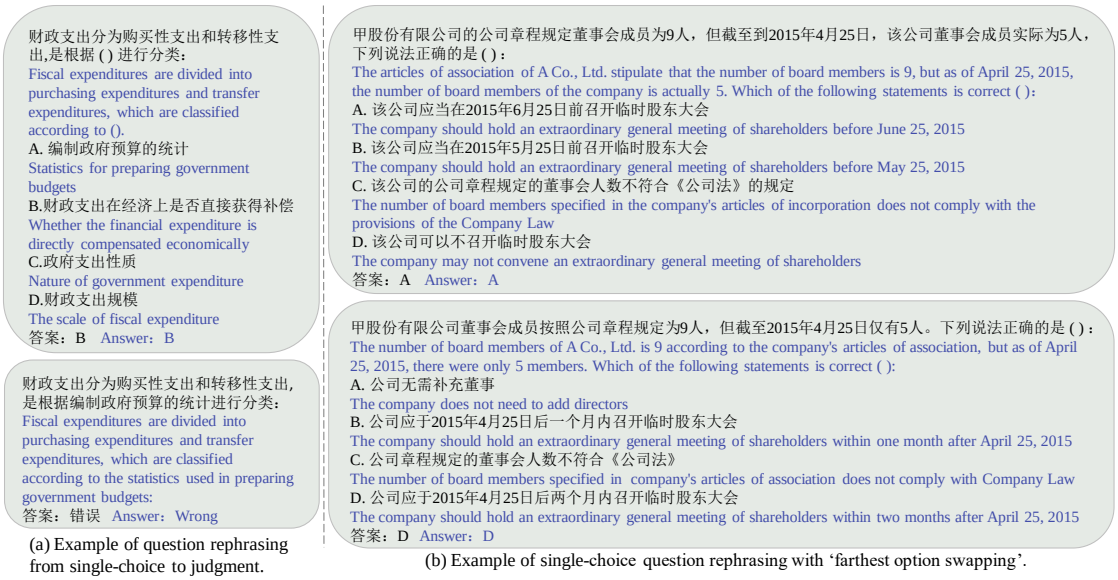
Following predecessors (Yuan et al., 2021; Penedo et al., 2023; Wei et al., 2023; Huang et al., 2024), all the collected questions go through a standard data preprocessing pipeline including cleaning and de-duplication. For data cleaning, we first remove non-Chinese paragraphs with the inexpensive n-gram models like fastText (Joulin et al., 2016). Then a series of filtering rules and heuristics are performed, such as only keeping lines with valid punctuation, discarding consecutive newlines and whitespace characters, or removing unsemantic and garbled lines. For data de-duplication, we adopt MinHash algorithm (Broder, 1997) for internal de-duplication and de-duplication with external public data (Zhang et al., 2023a; Zhang and Yang, 2023; Group, 2023; Information, 2023; Lu et al., 2023; Zhu et al., 2024).

To enhance data diversity and mitigate data contamination problem, we also adopt the strategy of question rephrasing based on GPT4 (Wang et al., 2022; Xu et al., 2023). We observe that the collected raw data exhibits a significant class imbalance, with a notable scarcity of judgment questions and a substantial surplus of single-choice questions. To address this issue, we prompt GPT4 to rephrase a portion of the single-choice questions into judgment questions, while maintaining semantic consistency, as exemplified in Figure 3 (a). Furthermore, to mitigate the problem of data contamination, we first randomize the option order (Berglund et al., 2023). In practice, this includes both random shuffle and 'farthest option swapping', where the correct option is exchanged with the incorrect option that is farthest away. Subsequently, we prompt GPT4 to rephrase the questions based on the shuffled options, similarly preserving semantic consistency, as exemplified in Figure 3 (b).

3.3.3 Human Cross-Validation

In order to ensure the quality of the benchmark, we establish a financial team of more than 10 people with professional financial backgrounds and conduct three rounds of manual verification of the rephrased questions. Specifically, we first spend about 2 months on the first round of verification on the rephrased questions. During this period,

[1] https://github.com/pymupdf/PyMuPDF
[2] https://pypi.org/project/EbookLib/



(a) Example of question rephrasing from single-choice to judgment.

(b) Example of single-choice question rephrasing with 'farthest option swapping'.

Figure 3: Examples of question rephrasing. English translations are shown in blue for better readability. In each example, the top is the original question, and the bottom is the rephrased question.

about 20,000 questions with obvious quality are discarded. About 70,000 questions with high quality and high confidence in the correctness of the answers are retained. In addition, about 35,000 questions for which it is difficult to determine the correctness of the answers in a short time are reserved for the next round of verification. In the second round of verification, we spend about another month carefully proofreading and verifying the questions, and produce about 30,000 high-quality questions. Finally, the final approximately 100,000 questions undergo a third round of random sampling and verification over 10 days to further improve the quality. The statistics of the final dataset are summarized in Appendix D.

4 Experiments

4.1 Setup

Data Split We randomly split the questions into a development set, a validation set, and a test set within each second-level category. The development split per category consists of three examples to facilitate few-shot evaluation. A portion of the development examples are also annotated with detailed explanations to enable few-shot chain-of-thought settings (Wei et al., 2022). The validation set and test set are divided in a ratio of 2:8, where the validation set is for hyperparameter tuning and the test set is for full evaluation.

Inference Details We employ the OpenCompass (Contributors, 2023) framework to perform model inference. Specifically, during the generation process, we set both the temperature and the top p to 1.0, and employ greedy decoding. The input token length is limited to 2048, and the output token length is limited to 64, which is sufficient for the questions of choice and judgment. Right truncation is performed for input prompts exceeding the length limitation. All models are inferred in both zero-shot and three-shot settings, which are exemplified in Appendix C.

Evaluation Metrics We adopt accuracy to measure the match between model prediction and gold answer. Specifically, for single-choice questions, if multiple valid options are predicted by the model, we only select the first option as the final answer predicted by the model. For multiple-choice questions, if any of the options predicted by the model are not among the gold answer, we directly classify it as wrong. Otherwise, we score it based on the number of predicted answers (out of a full score of 1). Considering that the judgment questions are relatively simple and usually score higher, we appropriately reduce the weight of the judgment questions to calculate the final score more reasonably, *i.e.*, $final = 0.4 \times single + 0.4 \times multiple + 0.2 \times judgment$.

4.2 Models

To give a comprehensive view of the status of LLMs in a Chinese financial context, we evaluate a wide spectrum of large language models, as depicted in Table 2. Specifically, open-source LLMs from various families are included, includ-

Table 2: The accuracy (%) on the test split under the answer-only setting. ⋆ represents the finance-specific LLMs.

Model	Subject		Qualification		Practice		Law		Average	
	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot
API Call										
GPT4	58.62	58.47	58.34	57.09	57.56	56.37	55.20	55.15	57.43	56.77
ChatGPT	39.58	40.86	43.15	42.56	40.86	40.51	38.18	38.83	40.44	40.69
Size > 65B										
Qwen2.5-72B	66.32	67.17	64.67	65.44	65.74	66.91	63.25	64.57	65.00	66.02
XuanYuan2-70B-Base⋆	53.56	59.96	53.36	56.27	53.55	58.05	48.27	52.48	52.19	56.69
Llama3-70B	50.75	56.27	47.52	52.35	45.17	51.24	44.62	49.26	47.02	52.28
DeepSeek-67B-Base	45.76	52.03	44.61	50.57	43.95	49.16	42.87	47.01	44.30	49.69
Tigerbot-70B-Base	43.22	52.13	43.15	48.42	40.32	46.00	38.56	45.87	41.31	48.11
Size ≈ 30B										
Qwen2.5-32B	61.77	63.31	62.03	64.44	60.42	62.25	58.61	61.17	60.71	62.79
Yi1.5-34B	58.62	61.44	58.30	60.91	57.26	59.75	55.16	58.55	57.34	60.16
10B < Size < 20B										
Qwen2.5-14B	56.18	59.55	58.37	62.08	56.64	60.13	55.58	58.71	56.69	60.12
InternLM2.5-20B	52.41	54.97	54.74	57.59	53.07	56.42	51.24	54.29	52.87	55.82
XuanYuan-13B-Base⋆	40.87	44.32	41.64	47.68	42.30	46.56	41.73	45.75	41.64	46.08
Phi3-14B-Instruct	44.95	42.46	46.12	43.60	44.26	40.83	42.17	39.61	44.38	41.63
Baichuan2-13B	29.16	41.13	34.25	44.19	31.27	40.63	31.44	40.05	31.53	41.50
Skywork-13B	34.66	39.24	37.78	43.88	36.22	40.90	36.39	41.38	36.26	41.35
DISC-FinLLM-13B⋆	37.42	39.46	42.59	38.81	38.32	39.89	39.09	38.35	39.36	39.13
Tigerbot-13B-Base	32.83	34.09	35.36	39.21	33.09	35.87	33.75	35.52	33.76	36.17
Llama2-13B	30.24	32.15	31.42	35.18	29.35	31.92	29.48	34.33	30.12	33.40
5B < Size < 10B										
Qwen2.5-7B	56.27	58.41	54.13	56.28	55.25	58.47	53.53	56.26	54.80	57.36
YunShan-7B⋆	52.65	53.00	52.61	51.79	53.33	52.77	52.52	52.23	52.78	52.45
InternLM2.5-7B	52.31	54.67	50.26	53.22	51.49	53.85	49.55	51.62	50.90	53.34
Yi1.5-9B	47.85	50.51	50.45	51.03	48.26	49.20	46.41	47.03	48.24	49.44
Qwen1.5-7B	46.26	49.18	47.97	50.28	46.52	47.93	44.66	46.04	46.35	48.36
ChatGLM3-6B-Base	46.56	46.41	47.52	49.45	46.56	48.20	43.62	45.05	46.07	47.28
XuanYuan-6B-Base⋆	39.99	41.65	44.30	45.87	42.70	43.91	41.70	42.81	42.17	43.56
Llama3.1-8B	33.54	40.38	34.87	41.25	32.54	39.47	35.24	41.27	34.05	40.59
Baichuan2-7B	30.22	37.23	35.56	41.35	31.30	37.33	29.59	37.49	31.67	38.35
Mistral-7B	29.46	35.63	29.11	37.56	28.75	35.87	28.39	34.34	28.93	35.85
CFGPT1-sft-7B-Full⋆	32.45	33.73	34.41	33.92	32.80	32.93	33.48	35.05	33.29	33.91
FinMA-7B⋆	23.71	22.74	24.92	25.86	22.71	20.71	22.34	23.52	23.42	23.21
Size < 5B										
Qwen2.5-3B	40.15	44.53	46.20	47.01	42.52	43.87	42.23	44.81	42.78	45.06
Phi3.5-3.8B-Instruct	41.37	40.50	43.48	44.12	40.75	40.98	40.80	42.77	41.60	42.09
YunShan-1.5B⋆	35.06	37.36	39.19	41.61	38.17	39.95	38.08	40.32	37.63	39.81
InternLM2.5-1.8B	30.60	35.12	36.81	37.66	31.02	32.59	33.56	34.18	33.00	34.89
Gemma2-2B	22.16	30.98	22.78	32.58	21.83	29.99	23.48	30.97	22.56	31.13

ing Llama (Touvron et al., 2023b; Meta, 2024), Qwen (Bai et al., 2023), ChatGLM (Zeng et al., 2022; Du et al., 2022), Baichuan (Yang et al., 2023a), InternLM (Team, 2023; Cai et al., 2024), Phi (Gunasekar et al., 2023; Li et al., 2023c; Abdin et al., 2024), DeepSeek (DeepSeek-AI, 2024), XuanYuan (Zhang and Yang, 2023), FinMA (Xie et al., 2023), Gemma (Team et al., 2024), TigerBot (Chen et al., 2023b), Skywork (Wei et al., 2023), Yi (Young et al., 2024), Mistral (Jiang et al., 2023), DISC-FinLLM (Chen et al., 2023a) and CFGPT (Li et al., 2023b). We classify models into different categories according to their size. Considering the legal issues, we only report the results of two API-based LLMs, *i.e.*, ChatGPT (*gpt-3.5-turbo-0125*) and GPT4 (*gpt-4-turbo*).

4.3 Results

In Table 2, we report the 0-shot and 3-shot accuracy of each first-level category on the test split under the answer-only setting. As can be seen, the Chinese-oriented Qwen2.5-72B (Bai et al., 2023) lead the benchmark, with an average accuracy reaching 66.02%. In addition, the accuracy of some

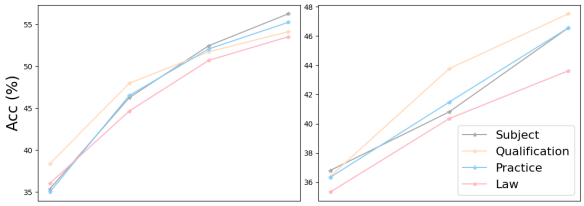
other models like Yi1.5-34B (Young et al., 2024), XuanYuan2-70B-Base (Zhang and Yang, 2023) and GPT4 (OpenAI, 2023) also exceed 56%. In the size range of 10B-20B, Qwen2.5-14B (Bai et al., 2023), InternLM2.5-20B (Cai et al., 2024) and XuanYuan-13B-Base (Zhang and Yang, 2023) are in the lead, with average accuracy exceeding 46%. Notably, in the size range of 5B-10B, Qwen2.5-7B (Bai et al., 2023) and the Chinese finance-specific model YunShan-7B (Wang et al., 2023) is in the absolute leading position, with an accuracy of over 52%, which is even higher than the accuracy of some 70B models. Also, Yi1.5-9B (Young et al., 2024), Qwen1.5-7B (Bai et al., 2023) and ChatGLM3-6B-Base (Zeng et al., 2022; Du et al., 2022) also achieve the accuracy over 45%. During the size range of less than 5B, Qwen2.5-3B (Bai et al., 2023) is the only model that achieves an accuracy of more than 45%. The other models like Phi3.5-3.8B-Instruct (Abdin et al., 2024) and YunShan-1.5B (Wang et al., 2023) also achieve the accuracies of more than 39%. In conclusion, there is still significant room for improvement.

4.4 Analysis

Few-shot examples are helpful in most cases. As we can see from Table 2, the performance of most models demonstrates improvement when some examples are provided. However, in the case of Phi3-14B-Instruct, the zero-shot setting outperforms the few-shot setting. We guess that this is because the models have acquired the ability to fully understand the questions without the need for examples during the pre-training or fine-tuning. The introduced examples may mismatch with their training methodology, which leads to the decrease in accuracy (Gu et al., 2024; Li et al., 2023a).

Scaling up the model size usually results in better performance. In Table 2, as the model size increases, the accuracy of Qwen2.5-3B, Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B and Qwen2.5-72B increases accordingly, *i.e.*, 45.06%, 57.36%, 60.12%, 62.79% and 66.02% respectively at the 3-shot setting. Similarly, the accuracy of Yi1.5-34B is increased by 10.72% over Yi1.5-9B at the 3-shot setting. However, this does not mean that increasing the model size will definitely improve the performance.

Domain specific pre-training and fine-tuning are helpful. Impressively, two finance-specific models XuanYuan (Zhang and Yang, 2023) and YunShan (Wang et al., 2023) achieve very competitive accuracy, demonstrating a better grasp of financial knowledge. This can be attributed to the fact that both models are mixed with high-quality financial corpus during pre-training and fine-tuning. We think the same is true for the general base models such as Yi (Young et al., 2024) and Qwen (Bai et al., 2023) that perform well in Table 2.



(a) Qwen-7b series (b) ChatGLM-6b series

Figure 4: The 0-shot accuracy changes with the iterations of LLMs.

Accuracy improves with LLMs' iteration. We visualize how the accuracies of 4 first-level categories change as the LLM is iteratively updated (Qwen, Qwen1.5, Qwen2, Qwen2.5 and ChatGLM, ChatGLM2, ChatGLM3). As can be seen from Figure 4, with the improvement of training method and the enrichment of training corpus, the accuracy will often be significantly improved, which also highlights the rationality and importance of CFinBench.

The performance of chat models. To better engage in natural conversation, the chat version are often derived from base model by alignment techniques (Ouyang et al., 2022), such as supervised finetuning (SFT) and reinforcement learning from human feedback (RLHF). As observed in Table 3, the accuracy of some models' chat version is improved when compared to the base version, such as Qwen2.5-32B, Gemma2-2B, Baichuan2-13B, *etc*. At the same time, the accuracy of some models' chat version have declined, such as Yi1.5-34B and ChatGLM3-6B. The varying alignment strategies of the models lead to different results.

Table 3: The 0/3-shot average accuracy (%) of base model and chat model on the test split.

Model	Base		Chat	
	0-shot	3-shot	0-shot	3-shot
Yi1.5-34B	57.34	60.16	58.99	57.48
Qwen2.5-32B	60.71	62.79	61.62	64.48
InternLM2.5-20B	52.87	55.82	54.35	55.27
Baichuan2-13B	31.53	41.50	40.74	44.60
ChatGLM3-6B	46.07	47.28	30.79	27.27
Gemma2-2B	22.56	31.13	30.47	35.81

5 Conclusion

In this paper, we present CFinBench, the most comprehensive evaluation benchmark to date, for assessing the financial domain knowledge of LLMs under Chinese context. We improve the quality and diversity of the data and mitigate the issue of data contamination through a series of processes, including data cleaning, internal and external deduplication, LLM-assisted question rephrasing, option shuffling, and multiple rounds of human cross-validation. Four first-level categories are included in CFinBench: financial subject, financial qualification, financial practice, and financial law, which are more align with the career trajectory of financial practitioners in China. The CFinBench comprises 99,100 questions spanning 43 second-level categories with 3 question types: single-choice, multiple-choice and judgment. We conduct extensive evaluations on a wide spectrum of LLMs with various model size. The results show that there is still significant room for improvement for current LLMs in the Chinese financial domain.

6 Ethical Statement

All data utilized in this study primarily originate from publicly accessible channels that have been processed by our professional experts. Also, It is important to note that all datasets within CFinBench carry low ethical risks, with stringent measures in place to ensure the absence of any sensitive or personally identifiable information. In order to promote the research progress, our data and evaluation code are open sourced to the community under the Apache-2.0 license[3].

7 Limitations

CFinBench focuses on Chinese financial system and is therefore not suitable for assessing financial knowledge in other countries. It provides an important basis for evaluating the LLMs' mastery of Chinese financial knowledge. However, since we have open sourced the questions and answers of the validation split, if they are improperly used to train the large language model, the accuracy of the model may be falsely high on the validation split. Therefore, the results on benchmark are just a reference. The true quality of the model depends on the performance of the user in the practical scenario.

[3]https://www.apache.org/licenses/LICENSE-2.0

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023a. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023b. Tigerbot: An open multilingual multitask llm. *arXiv preprint arXiv:2312.08688*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

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*.

Ant Group. 2023. Ant-fin-eva. https://github.com/alipay/financial_evaluation_dataset/.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, et al. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18099–18107.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Gang Hu, Ke Qin, Chenhan Yuan, Min Peng, Alejandro Lopez-Lira, Benyou Wang, Sophia Ananiadou, Wanlong Yu, Jimin Huang, and Qianqian Xie. 2024. No language is an island: Unifying chinese and english in financial large language models, instruction data, and benchmarks. *arXiv preprint arXiv:2403.06249*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

East Money Information. 2023. Openfindata. https://github.com/open-compass/OpenFinData/.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yang Lei, Jiangtong Li, Ming Jiang, Junjie Hu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*.

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

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023b. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

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

Meta. 2024. Llama3. Https://llama.meta.com/llama3.

OpenAI. 2021. Gpt-3.5-turbo. Https://www.openai.com/chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Raj Sanjay 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 pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Robert J Shiller. 2013. *Finance and the good society*. Princeton University Press.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 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 models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yunhe Wang, Hanting Chen, Yehui Tang, Tianyu Guo, Kai Han, Ying Nie, Xutao Wang, Hailin Hu, Zheyuan Bai, Yun Wang, et al. 2023. Pangu-π: Enhancing language model architectures via nonlinearity compensation. *arXiv preprint arXiv:2312.17276*.

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.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

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. *arXiv preprint arXiv:2306.05443*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023b. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*, 2:65–68.

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, et al. 2023a. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Xuanyu Zhang, Bingbing Li, and Qing Yang. 2023b. Cgce: A chinese generative chat evaluation benchmark for general and financial domains. *arXiv preprint arXiv:2305.14471*.

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.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Jie Zhu, Junhui Li, Yalong Wen, and Lifan Guo. 2024. Benchmarking large language models on cflue–a chinese financial language understanding evaluation dataset. *arXiv preprint arXiv:2405.10542*.

A More experiments

Results on chain-of-thought setting. To further explore the models' reasoning capabilities, in addition to the answer-only (AO) setting, we also perform some experiments on the chain-of-thought (COT) setting (Kojima et al., 2022; Wei et al., 2022). Evaluation on COT setting requires the model to generate explanations for a given question and then give the final answer based on the generated explanations. Specifically, we obtain the explanations examples only for multiple-choice questions manually by professional financial practitioners. The experimental results are reported in Table 4.

As observed in Table 4, the models achieve comparable or lower average accuracy than in the answer-only setting. This suggests that COT prompting does not necessarily improve results, which is also evidenced in other benchmarks like FinEval (Zhang et al., 2023a) and C-Eval (Huang et al., 2024), *etc*.

Comparison with other similar benchmark. FinEval (Zhang et al., 2023a) is another representative benchmark for evaluating the Chinese financial advanced knowledge of LLMs. In Table 5, we report the few-shot accuracy on CFinBench and FinEval with various LLMs. It can be seen that the accuracy of the highest Yi1.5-34B reaches nearly 87%, while ours is around 60%. Likewise, the lowest InternLM2-7B accuracy is still over 62%, while ours is only around 43%. This all suggests that CFinBench is more challenging and better able to distinguish the performance of different models.

B Analysis on Data Contamination

Since we publicize the answers of the validation set, we imagine a data contamination scenario: what would happen if the LLM was trained using part of the validation set? We conduct an experiment here. Specifically, we randomly select 60% of the data from the validation set and fine-tuned InternLM2.5-1.8B for 2 epochs based on LoRA (the modules of query, key, value and out are chosen as the target, the rank is set to 32). The 0-shot results on validation and test split are reported in Table 6. As we can see, data contamination on validation split will lead to the accuracy improvements on both validation and test. Therefore, it is crucial to avoid the contamination of evaluation data during the model training process.

C Prompt examples

We list the prompt examples utilized in the evaluation process, including zero-shot and few-shot in answer-only scenarios (Figure5 and Figure6), zero-shot and few-shot in chain-of-thought scenarios (Figure7).

D Data statistic

In Table 7, we enumerate the comprehensive statistical information of the dataset.

Table 4: The 0-shot and 3-shot average accuracy (%) of multiple-choice questions on the test split under the answer-only (AO) setting and chain-of-thought (COT) setting.

	GPT4		Qwen-14B		ChatGLM3-6B-Base	
	0-shot	3-shot	0-shot	3-shot	0-shot	3-shot
AO	48.52	48.67	36.51	40.66	37.27	37.58
COT	49.23	48.95	27.65	29.96	29.48	31.11

Table 5: The few-shot average accuracy (%) of CFinBench and FinEval under the answer-only setting.

	GPT4	Yi1.5-34B	Qwen1.5-72B	Qwen1.5-32B	Qwen1.5-7B	ChatGLM3-6B	InternLM2-7B
CFinBench	56.77	60.16	58.56	57.64	48.36	47.28	43.65
FinEval	70.13	86.79	83.93	84.36	72.55	63.08	62.03

Table 6: The 0-shot accuracy (%) of InternLM2.5-1.8B on validation and test split with or without data contamination.

	Subject	Qualification	Practice	Law	Average
Val (W/O)	32.52	35.43	33.71	34.24	33.98
Test (W/O)	30.60	36.81	31.02	33.56	33.00
Val (W)	75.13	82.79	76.22	80.15	78.57
Test (W)	40.62	47.21	40.89	42.68	42.85

请你做一道关于金融学的单项选择题，你将从A，B，C，D中选出正确的答案，并写在‘答案：’之后，答案应只包含最终结果，不要添加额外语句。例如：答案：B。请你严格按照上述格式作答。题目如下：
You are asked to do a multiple choice question on Finance, you will select the correct answer from A, B, C, D and write it after ‘Answer:’ The answer should contain only the final result without adding extra statements. For example, Answer : B. Please follow the above format strictly. The questions are as follows:
广义外汇与狭义外汇最本质的差异是什么？
What is the most essential difference between broad and narrow foreign exchange?
A、 能否用于偿还债务 B、 是否能够在国际上自由兑换 C、 是否可用本国货币表示 D、 是否涵盖所有有价证券
A、What is the most essential difference between broad and narrow foreign exchange? B、 Availability for debt servicing C、 Can be expressed in national currency D、 Whether all marketable securities are covered
答案： Answer:

(a) Example of single-choice questions on Finance.

请你做一道关于税务师的多项选择题，答案可能是一个到多个选项,请你从四个或者五个选项中选出正确的答案，并将其写在'答案：'之后，答案应只包含最终结果，不要添加额外语句。例如：答案: A D。请你严格按照上述格式作答。题目如下：
You are asked to do a multiple choice question on Tax Accountant. The answer may be one to more than one option. Please select the correct answer from four or five options and write it after ‘Answer:’. The answer should contain only the final result without adding extra statements. For example: Answer: A D. Please follow the above format strictly. The questions are as follows:
关于税务咨询，下列哪项说法是正确的( )?
Which of the following statements is true about tax counselling ( )?
A、面对面回答纳税人的税务疑问被视为口头咨询 B、利用电话进行的咨询可以被称作口头咨询 C、通过电子邮件形式进行的咨询被划分为网络咨询 D、网络咨询代表了税务咨询方式的一种新趋势 E、书面咨询是最普遍采用的税务咨询方法
A、 Face-to-face answers to taxpayers‘ tax queries are treated as verbal advice B、 Counselling using the telephone can be referred to as verbal counselling C、 Counselling via email is classified as web counselling D、 Web-based counselling represents a new trend in tax counselling methods E、 Written advice is the most commonly used method of tax advice
答案： Answer:

(b) Example of multiple-choice questions on Tax Accountant.

请回答下面关于银行业法的判断题，将你的判断结果写在'答案：'之后，答案应只包含最终结果，不要添加额外语句。若给定题目表述正确，则回答：'答案：正确'，否则回答：'答案：错误'。请严格按照上述格式作答。题目如下：
Answer the following judgement question on Banking Law by writing your judgement after ‘Answer:’ The answer should contain only the final result without adding extra statements. If the given question is correctly stated, answer: ‘Answer: correct’, otherwise answer: ‘Answer: incorrect’. Please follow the above format strictly. The questions are given below:
商业银行的合规经营是基于所有从业人员的合规行为建立的，因此，即使是个别银行从业人员的违规行为也不会对银行整体的合规性产生影响。
The compliance of a commercial bank is established on the basis of the compliant behavior of all practitioners, and therefore, even the non-compliance of an individual banker will not have an impact on the compliance of the bank as a whole.
答案： Answer:

(c) Example of judgment questions on Banking Law.

Figure 5: Examples of zero-shot prompts in answer-only setting. English translations are shown in blue for better readability.

请你做一道关于财政学的单项选择题，你将从A，B，C，D中选出正确的答案，并写在'答案：'之后，答案应只包含最终结果，不要添加额外语句。例如：答案： B。请你严格按照上述格式作答。下面会展示三个问题及其答案的示例。题目如下：
You are asked to do a single choice question on Public Finance, you will select the correct answer from A, B, C, D and write it after 'Answer:' The answer should contain only the final result without adding extra statements. For example, Answer : B. Please follow the above format strictly. Samples of three questions and their answers will be shown below. The questions are as follows:
我国的中央税包括以下哪一种税种？
Central taxes in the country include which of the following?
A、 企业所得税 B、 营业税 C、 消费税 D、 增值税 答案: C
A. Enterprise income tax B. Business tax C. Consumption tax D. Value-added tax Answer: C

…[3-shot examples]…

从狭义的角度出发，被认为有权限制定影响税收分配过程中权利义务关系的法律规范的国家机关是指什么？
In a narrow sense, what is meant by the State organ considered to have competence to establish legal norms affecting the relations of rights and obligations in the process of tax allocation?
A、中华人民共和国国务院 B、中国国税总局 C、中央财政部 D、全国人大及其常委会 答案：
A. State Council of the People's Republic of China B. China's State Administration of Taxation C. Central Ministry of Finance
D. National People's Congress and its Standing Committee Answer：

(a) Example of single-choice questions on Public Finance.

请你做一道关于政治经济学的多项选择题，答案可能是一个到多个选项,请你从四个或者五个选项中选出正确的答案，并将其写在'答案：'之后，答案应只包含最终结果，不要添加额外语句。例如：答案: A D。请你严格按照上述格式作答。下面会展示三个问题及其答案的示例。题目如下：
如果小王用38元在店里购买了一副网球拍，而店家需要将8元缴纳作为税费，那么在此交易中，货币的作用是什么？
Please do a comparison question about political economy. The answer may be four or more options. Please choose the correct answer from one or five options and write it after 'Answer:'. The answer to the question should be only Contain the final result and do not add additional statements. For example: Answer: A D. Please answer strictly according to the above format. Examples of three questions and their answers are shown below. as follows:
If Wang buys a tennis racket in the store for 38 yuan, and the store needs to pay a fine of 8 yuan as tax, then what is the role of currency in the transaction?
A、 价值尺度 B、 流通手段 C、 支付手段 D、 贮藏手段 答案: B, C
A. Value scale B. Circulation means C. Payment means D. Storage means Answer: B、C

…[3-shot examples]…

在市场经济环境中，企业作为市场活动的基本和关键参与者，其特点包括( )
In a market economy environment, enterprises serve as basic and key participants in market activities, and their characteristics include ()
A、企业须独立进行财务核算，并承担盈亏责任 B、企业生产产品或服务的首要目标是获得收益 C、企业负责直接向社会供应产品及服务
D、企业具备自主经营的权利 答案：
A. Enterprises must conduct financial accounting independently and assume responsibility for profits and losses. B. The primary goal of enterprises in producing products or services is to obtain profits. C. Enterprises are responsible for directly supplying products and services to the society. D. Enterprises have the right to operate independently. Answer：

(b) Example of multiple-choice questions on Political Economy.

请回答下面关于保险从业资格的判断题，将你的判断结果写在'答案：'之后，答案应只包含最终结果，不要添加额外语句。若给定题目表述正确，则回答：'答案：正确'，否则回答：'答案：错误'。请严格按照上述格式作答。下面会展示三个问题及其答案的示例。题目如下：
Please answer the following judgment questions about insurance qualification. Write your judgment results after 'Answer:'. The answer should only contain the final result and do not add additional sentences. If the given question is stated correctly, answer: 'Answer: Correct', otherwise answer: 'Answer: Wrong'. Please answer strictly according to the above format. Examples of three questions and their answers are shown below. The questions are as follows:
车辆损失可以全额获得车损险赔偿，无论是否存在过错。
Vehicle damage insurance is fully reimbursable for vehicle damage, regardless of fault.
答案: 正确 Answer：Correct

…[3-shot examples]…

保险公司的理赔必须在所有情况下满足客户的要求。
The insurance company's claims must meet the customer's requirements in all circumstances.
答案： Answer：

(c) Example of judgment questions on Insurance Qualification.

Figure 6: Examples of few-shot prompts in answer-only setting. English translations are shown in blue for better readability.

请你做一道关于政治经济学的多项选择题，我们将在下面提供3个示例供参考，首先你需要对问题作一步一步地思考，答案可能是一个到多个选项，请你从四个或者五个选项中选出正确的答案，并在思考后将其写在'答案：'之后。以下为三个问题及其思考过程和最终答案的示例。题目如下：
Take a multiple-choice question about political economy. We will provide 3 examples below for reference. First, you need to think about the question step by step. The answer may be one to multiple options. Please choose from four Or choose the correct answer from the five options and write it after 'Answer:' after thinking about it. Below are examples of three questions, their thought processes, and final answers. The questions are as follows:
近期，中国东南部沿海区域经历了广泛的“民工短缺”现象，而农村地带却有一定数量的劳动力过剩。这种情况表明了什么？
Recently, China’s southeastern coastal areas have experienced a widespread "migrant worker shortage" phenomenon, while there is a certain amount of labor surplus in rural areas. What does this situation indicate?
A、 应结合国家的宏观调控和市场调节机制共同作用来解决这一问题 B、 市场在调节中存在不可预见性和延迟性，因此不能仅仅依靠市场机制来解决这一矛盾 C、 市场在分配资源方面扮演着根本性角色，这一矛盾应当完全交由市场调节来处理 D、 市场调控功能已经无法发挥，需通过计划经济手段来调整劳动力资源，以解决这一问题
A、This problem should be solved by combining the national macro-control and market adjustment mechanisms. B、The market has unpredictability and delay in adjustment, so this contradiction cannot be solved by relying solely on the market mechanism. C、The market is allocating resources. Playing a fundamental role in this aspect, this contradiction should be completely left to market regulation. D、The market regulation function can no longer be exerted, and labor resources need to be adjusted through planned economic means to solve this problem.
在以上选项中，应该选择什么？ Which of the above options should you choose?
让我们逐步思考：市场经济是社会化的商品经济，是市场在资源配置中起基础作用的经济，出现这种劳动力不平衡的状况，是由于市场自发调节具有盲目性和滞后性的缺点，要解决这一问题，一方面要发挥市场的调节作用，另一方面要依靠国家宏观调控，弥补市场调节的不足。因此正确选项为A, B。答案: A, B
Let us think step by step: The market economy is a socialized commodity economy, an economy in which the market plays a basic role in resource allocation. This imbalance in the labor force is due to the shortcomings of blindness and lag in spontaneous market regulation. It must be To solve this problem, on the one hand, we must give full play to the regulatory role of the market, and on the other hand, we must rely on national macro-control to make up for the lack of market regulation. Therefore the correct options are A and B. Answer: A, B

…[3-shot examples]…

产业政策在国家宏观调控政策中包含哪些部分？ What parts does industrial policy include in the national macro-control policy?
A、产业布局政策 B、产业组织政策 C、产业结构政策 D、产业优惠政策
A. Industrial layout policy B. Industrial organization policy C. Industrial structure policy D. Industrial preferential policy
在以上选项中，应该选择什么？让我们逐步思考:
Which of the above options should you choose? Let's think about it step by step:

(a) Few shot example of multiple-choice questions on Political Economy.

请你做一道关于期货从业资格的多项选择题，首先你需要对问题作一步一步地思考，然后请你从四个或者五个选项中选出正确的一个或多个答案，并将其写在'答案：'之后。题目如下：
Take a multiple-choice question about futures practitioner qualification. First, you need to think about the question step by step. Then you are asked to choose the correct answer or answers from four or five options and write them down. after 'Answer:'. The topics are as follows:
中国期货业协会采取了哪些措施来培育和提升期货投资咨询人员的专业水平？
What measures has the China Futures Association taken to cultivate and enhance the professional level of futures investment consulting personnel?
A、坚决打击侵害投资者利益、损害行业整体利益的违法失信行为，营造有利于期货公司进行投资咨询业务的环境 B、制定相关政策，规范期货从业人员的执业行为 C、借鉴国际及国内成熟市场经验，推动期货投资咨询资格认证体系的发展 D、强化期货从业人员的自我约束与管理
A. Resolutely crack down on illegal and untrustworthy behaviors that infringe on the interests of investors and harm the overall interests of the industry, and create an environment conducive to futures companies’ investment consulting services. B. Formulate relevant policies to standardize the professional behavior of futures practitioners. C. Learn from mature international and domestic markets. Experience, promote the development of futures investment consulting qualification certification system D. Strengthen the self-discipline and management of futures practitioners
在以上选项中，应该选择什么？让我们逐步思考：
Which of the above options should you choose? Let's think about it step by step:

(b) Zero shot example of multiple-choice questions on Futures Practitioner Qualification.

Figure 7: Examples of prompts in chain-of thought setting. English translations are shown in blue for better readability.

Table 7: The detailed statistic of CFinBench.

Category	Single-Choice	Multiple-Choice	Judgment	All
Subject	***3302***	***1889***	***3915***	***9106***
Political Economy (政治经济学)	115	47	67	229
Western Economics (西方经济学)	268	46	212	526
Microeconomics (微观经济学)	295	29	221	545
Macroeconomics (宏观经济学)	21	117	294	432
Industrial Economics (产业经济学)	406	199	249	854
Public Finance (财政学)	167	86	156	409
International Trade (国际贸易学)	99	48	100	247
Statistics (统计学)	974	794	1846	3614
Auditing (审计学)	443	429	381	1253
Economic History (经济史)	248	63	133	444
Finance (金融学)	266	31	256	553
Qualification	***14604***	***8879***	***5905***	***29388***
Tax Practitioner Qualification (税务从业资格)	1332	1544	464	3340
Futures Practitioner Qualification (期货从业资格)	2086	1396	1049	4531
Fund Practitioner Qualification (基金从业资格)	3892	118	536	4546
Real Estate Practitioner Qualification (地产从业资格)	503	511	660	1674
Insurance Practitioner Qualification (保险从业资格)	1780	1220	903	3903
Securities Practitioner Qualification (证券从业资格)	2734	2041	1518	6293
Banking Practitioner Qualification (银行从业资格)	258	173	266	697
Certified Public Accountant (CPA) (注册会计师)	2019	1876	509	4404
Practice	***18824***	***13419***	***9802***	***42045***
Junior Auditor (初级审计师)	317	317	194	828
Intermediate Auditor (中级审计师)	237	223	197	657
Junior Statistician (初级统计师)	158	190	97	445
Intermediate Statistician (中级统计师)	259	400	195	854
Junior Economist (初级经济师)	2262	1496	655	4413
Intermediate Economist (中级经济师)	2547	1250	913	4710
Junior Banking Professional (初级银行从业人员)	2886	2075	1646	6681
Intermediate Banking Professional (中级银行从业人员)	2572	1550	1482	5604
Junior Accountant (初级会计师)	1654	1217	964	3835
Intermediate Accountant (中级会计师)	1252	858	700	2810
Tax Consultant (税务师)	934	1115	493	2542
Asset Appraiser (资产评估师)	1779	1690	896	4365
Securities Analyst (证券分析师)	1967	1038	1370	4375
Law	***7695***	***5438***	***5428***	***18561***
Tax Law I (税法 I)	287	284	237	808
Tax Law II (税法 II)	283	323	238	844
Tax Inspection (税务稽查)	974	874	1664	3512
Commercial Law (商业法)	331	599	201	1131
Securities Law (证券法)	1009	106	693	1808
Insurance Law (保险法)	69	57	42	168
Economic Law (经济法)	610	424	405	1439
Banking Law (银行业法)	2783	1360	1231	5374
Futures Law (期货法)	922	884	477	2283
Financial Law (金融法)	315	323	180	818
Civil Law (民法)	112	204	60	376
Total	***44425***	***29625***	***25050***	***99100***