

SuperCLUE-Fin: Graded Fine-Grained Analysis of Chinese LLMs on Diverse Financial Tasks and Applications

Liang Xu, Lei Zhu, Yaotong Wu, Hang Xue

SuperCLUE team
contact@superclue.ai

Abstract

The SuperCLUE-Fin (SC-Fin) benchmark is a pioneering evaluation framework tailored for Chinese-native financial large language models (FLMs). It assesses FLMs across six financial application domains and twenty-five specialized tasks, encompassing theoretical knowledge and practical applications such as compliance, risk management, and investment analysis. Using multi-turn, open-ended conversations that mimic real-life scenarios, SC-Fin measures models on a range of criteria, including accurate financial understanding, logical reasoning, clarity, computational efficiency, business acumen, risk perception, and compliance with Chinese regulations.

In a rigorous evaluation involving over a thousand questions, SC-Fin identifies a performance hierarchy where domestic models like GLM-4 and MoonShot-v1-128k outperform others with an A-grade, highlighting the potential for further development in transforming theoretical knowledge into pragmatic financial solutions. This benchmark serves as a critical tool for refining FLMs in the Chinese context, directing improvements in financial knowledge databases, standardizing financial interpretations, and promoting models that prioritize compliance, risk management, and secure practices.

We create a contextually relevant and comprehensive benchmark that drives the development of AI in the Chinese financial sector. SC-Fin facilitates the advancement and responsible deployment of FLMs, offering valuable insights for enhancing model performance and usability for both individual and institutional users in the Chinese market.¹

1 Introduction

Science and technology finance as well as digital finance are the development trend of The Times when science and technology penetrate into the financial field. The research and development of large financial model is in line with the characteristics of this era and the development needs of enterprises, and provides a new technology enabling idea for the financial industry to better serve the society.

The penetration and integration of large language models into various industries is a technological development trend. As far as the financial industry is concerned, many

large financial models have gradually emerged in this era. Therefore, it has become an important and necessary topic how to provide accurate quantitative evaluation criteria for the Chinese native financial large model in strict accordance with the financial supervision system and combined with the knowledge in the financial field, and timely feedback on the development of the large model.

In order to evaluate the level of development of big financial models and make suggestions for improvement, we have released the SuperCLUE-Fin (SC-Fin) Chinese native big financial Model evaluation benchmark. According to different task types, the financial model is evaluated in an all-round and multi-angle way.

2 SuperCLUE-Fin

Characteristics

1) Assessment of Chinese native financial ability Based on providing Chinese world evaluation infrastructure for the era of general artificial intelligence, text input or prompt words are native Chinese; Taking full account of the development status of China's financial industry and the characteristics of China's financial regulatory system, we are committed to creating financial model evaluation indicators suitable for the Chinese context.

2) Wide range of financial application scenarios This evaluation sets up six application scenarios related to the financial field and 25 subdivided task types, covering various practical problems from macro to micro, from abstract to specific in various industries in the financial field, aiming to investigate the comprehensive strength of the financial model in dealing with the above various tasks from an all-round and multi-perspective.

3) Various ways of investigating problems This evaluation adopts a variety of question investigation methods, in form is divided into a single round of questions and answers, multiple rounds of questions and answers; The question types include noun explanation, calculation, brief answer, material analysis, comprehensive demonstration, etc., aiming to investigate the ability of the financial model to deal with various application problems from brief answer to complex through diversified questioning methods.

4) Open multi-round interactive QA In order to fully restore the user experience scene, most of the questions in this

Preprint.

¹Our benchmark can be found at <https://www.CLUEbenchmarks.com>

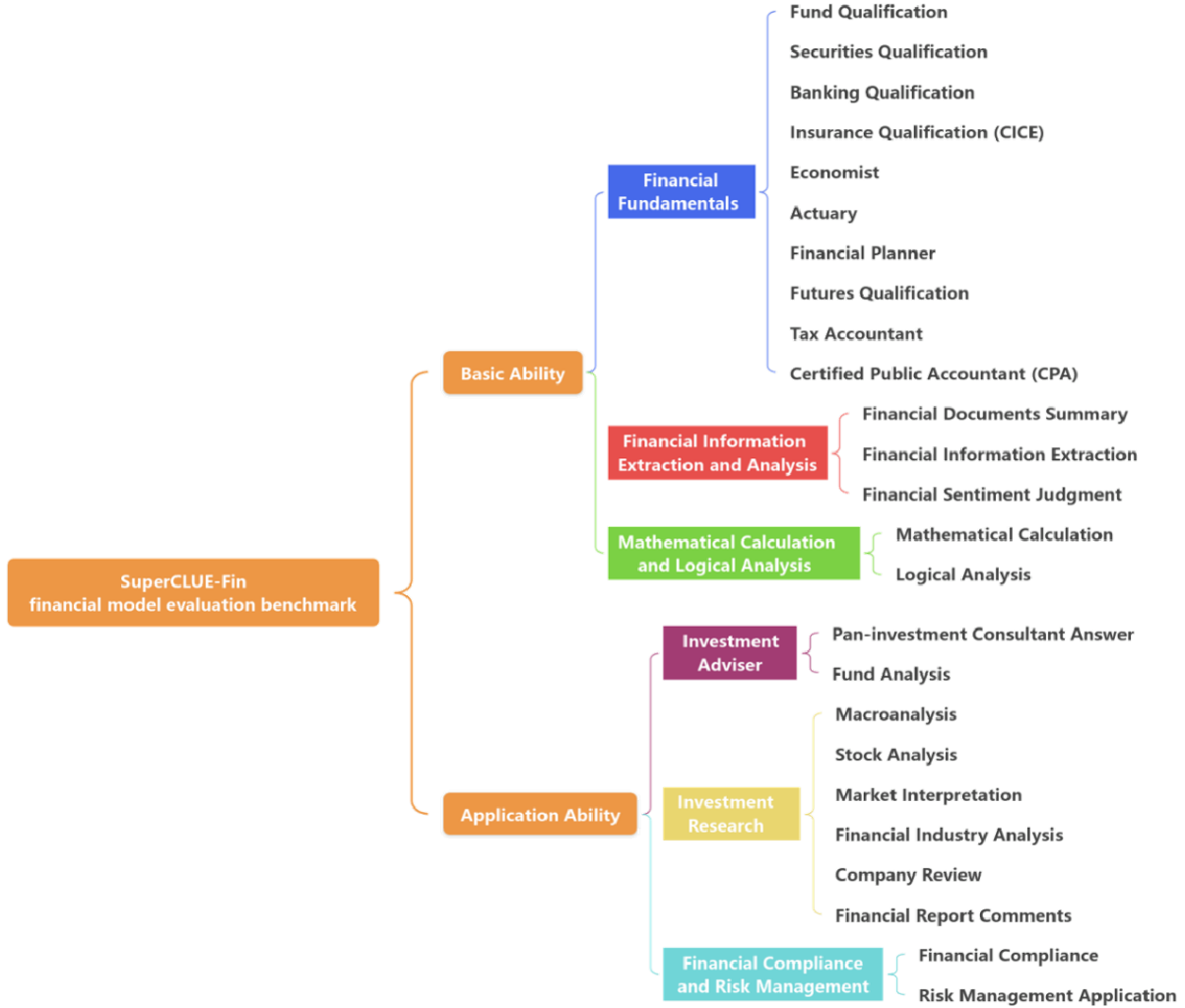


Figure 1: Overview of SuperCLUE-Fin task

evaluation adopt open multi-round interactive QA (a few are single round), and no objective questions are used as evaluation data. The purpose is to highly reflect the interactive experience between users and the large financial model, and make a more comprehensive and real investigation on the interactive QA ability of the model.

Dataset and Task Dimensions

The dataset contains over 1000 questions divided among numerous tasks, touching upon various financial sectors such as fund management, securities, futures, insurance, investment, wealth management, taxation, and banking. The tasks are classified into two primary dimensions:

1)Basic Capabilities: This includes tasks like Financial Fundamentals, Financial Information Extraction and Analysis, and Mathematical Calculation and Logical Analysis.

2)Applied Capabilities: This dimension focuses on tasks related to Financial Compliance and Risk Management, Investment Research, and Investment Adviser.

The content of each level task level sub-category is introduced as follows:

a. Financial Fundamentals(FF): examine the model’s mastery of basic knowledge in the financial field, including funds, securities, futures, insurance, investment, financial management, taxation, banking and so on. Specific sub-category tasks include: Fund Qualification, Securities Qualification, Banking Qualification, Insurance Qualification (CICE), Economist, Actuary, Financial Planner, Futures Qualification, Tax Accountant, Certified Public Accountant (CPA).

b. Financial Information Extraction and Analysis(FIEA): examine the model’s ability to interpret documents related to the financial field and its ability to judge and understand financial entities, financial intentions, financial emotions and other contents. Specific sub-tasks include: Financial Documents Summary, Financial Information Extraction, Financial Sentiment Judgment.

c. Mathematical Calculation and Logical Analysis(MCLA): examine the model’s ability to identify and

问题:

为什么基金管理人的内部环境对风险管理至关重要?

Why is the internal environment of fund managers critical to risk management?

模型答案:

基金管理人的内部环境对风险管理至关重要,因为它为基金风险管理提供了基础架构、指导原则和文化环境。内部环境是风险管理策略、流程和实践有效实施的前提条件。以下是几个关键点来解释这一重要性:

1. 制定风险管理政策和流程: 基金管理人的内部环境决定了其风险管理政策的制定方式以及如何实现这些政策。一个稳固的内部环境能够确保这些政策和流程得到有效的制定、执行和监督。

...<中间省略>...

6. 培养风险管理文化: 内部环境塑造了企业文化,包括对风险的态度和价值观。一个将风险管理视为核心价值的组织,能够更好地预防风险和应对可能出现的危机。

总之,基金管理人的内部环境对风险管理的重要性不容忽视。一个健全的内部环境能够提升风险管理的效率和效果,保护投资者利益,并维护市场的健康和稳定。

The internal environment of a fund manager is critical to risk management because it provides the infrastructure, guiding principles and cultural environment for risk management in fund management. The internal environment is a precondition for the effective implementation of risk management strategies, processes and practices. Here are a few key points to explain this importance:

1. Develop risk management policies and procedures: The internal environment of a fund manager determines how its risk management policies are developed and how they are implemented. A solid internal environment ensures that these policies and processes are effectively developed, implemented and monitored.

... < middle omission > ...

6. Cultivate a risk management culture: The internal environment shapes the corporate culture, including attitudes and values towards risk. An organization that regards risk management as a core value is better able to prevent risks and respond to possible crises.

In short, the importance of the internal environment of fund managers to risk management cannot be ignored. A sound internal environment can improve the efficiency and effectiveness of risk management, protect the interests of investors, and maintain the health and stability of the market.

评价结果:

我的思考与答案:

基金管理人的内部环境是风险管理的基石。一个健全和有效的内部环境可以促进风险管理政策的制定和执行,加强法规遵从性,以及培养积极的风险管理文化。内部环境涵盖组织结构、管理层的领导和支持、内部控制系统、员工的行为准则等多方面。无论是从防范系统性风险还是操作性风险的角度来看,内部环境的作用都是不可忽视的。它对于确保资金的安全性、流动性和盈利性起到核心作用,并且有助于维持公司的声誉和投资者信心。

问题及改进的反馈意见:

AI助手回答的质量比较高,覆盖了内部环境对风险管理的多个影响层面。然而,AI助手模型的回答缺少了专业术语或金融理论的特定引用,且没有提及具体的法规示例。更多的细节或实际案例将使答案更完整。

评分:

1. 基于符合基本要求:AI助手的回答涵盖了用户问题的核心主题,但可以通过提供具体例子来丰富和完善。根据标准,可得4分;

2. 基于知识正确:AI助理在风险管理相关知识准确,不提供错误的专业信息,基于标准的4分;

3. 基于语言与清晰度:答案的语言清晰,逻辑结构清晰,易于用户理解,基于标准的4分;

综合起来,人工智能助理的回答得到了4分的综合分数。

My thoughts and answers:

The internal environment of fund managers is the cornerstone of risk management. A sound and effective internal environment can facilitate the development and enforcement of risk management policies, enhance regulatory compliance, as well as foster a positive risk management culture. The internal environment covers many aspects such as organizational structure, leadership and support of management, internal control system, code of conduct of employees and so on. Whether from the perspective of preventing systemic risk or operational risk, the role of internal environment cannot be ignored. It plays a central role in ensuring the safety, liquidity, and profitability of funds, and helps maintain a company's reputation and investor confidence.

Problems and feedback on improvement:

The quality of the AI assistant's answers is relatively high, covering multiple levels of influence of the internal environment on risk management. However, the AI assistant model's answers were missing specific references to technical terms or financial theory and did not mention specific examples of regulations. More details or actual cases would make the answer more complete.

Start of scoring and evaluation:

1. Based on Basic Requirement: The AI assistant's reply covers the core theme of the user's question, but it can be enriched and improved by providing specific examples. Based on the standard, it can get 4 points;

2. Based on Information Correctness: the AI assistant is accurate in the relevant knowledge points of risk management and does not provide wrong professional information; 4 points based on criteria;

3. Based on Language Intelligibility: the language of the answer is clear, the logical structure is clear, and it is easy to be understood by users; 4 points based on criteria;

Taken together, the AI assistant's responses yielded a composite score of 4.

Figure 2: An example of a problem in SC-Fin on QA

logically analyze actual scenarios in the financial field and its ability to skillfully use financial formulas to solve practical problems. Specific sub-tasks include:Mathematical Calculation, Logical Analysis.

d. Financial Compliance and Risk Management(FCRM): examine the model's ability to master and apply domestic financial regulatory systems and industry standards, as well as its ability to assist users in analyzing potential financial risks of business and proposing corresponding control measures. Specific sub-tasks include: Financial Compliance, Risk Management Application.

e. Investment Research(IR): investigate the model's ability to control and analyze macro economy, market trend, industry development, corporate business and stock market conditions, and assist users to make correct judgments and decisions in the investment research stage. Specific sub-tasks include: Macroanalysis, Stock Analysis, Market Interpretation, Financial Industry Analysis, Company Review, Financial Report Comments.

f. Investment Adviser(IA): inspection model as the user's

investment consultant to assist users to complete the analysis and formulation of investment strategies. Specific sub-category tasks include: Pan-investment Consultant Answer, Fund Analysis.

Three categories of examples are provided as figure2-4.

Evaluation Dimensions

Each response is evaluated across multiple criteria, including Basic Requirement, Information Correctness, Logic of Finance, Language Intelligibility, Mathematical, Ability and Efficiency, Ability to Analyze Business, Clear Judgment, Safety Measure, Financial Security and Compliance, Risk Prediction and Control, Financial Insight. Scores are assigned on a five-point scale for each criterion, and the final model score is calculated as the average across all tasks. The definitions of the evaluation criteria are as follows:

a. Basic Requirement(BR): The answer should follow the user's intention, meet the basic purpose and needs of the user to ask the question appropriately.



Figure 3: An example of a problem in SC-Fin on the problem of calculation

b. Information Correctness(IC): The interpretation and use of professional words involved in the answer should be completely correct, including professional terms, financial formulas, financial theories, and related knowledge in the economic field.

c. Logic of Finance(LF): The analytical logic of the answer should have a certain degree of professionalism, and demonstrate advanced thinking and expertise in financial information interpretation is needed for the perspective of solving the problem .

d.Language Intelligibility(LI): The answers are clear and easy to understand, using concise language and expressions so that users can easily understand them.

e. Mathematical Ability and Efficiency(MAE): The mathematical calculation problem in the financial field should be responded to efficiently and give correct results. At the same time, it should have a detailed analysis and calculation process, and the formula reference should be correct.

f. Ability to Analyze Business(AAB): Professional analysis of business needs and business types, such as products,

markets, investments, finance, etc.

g. Clear Judgment(CJ): The answer should give a clear and directional judgment on the basic situation such as financial intention, financial sentiment, financial industry trend, and economic situation.

h. Safety Measure(SM): Financial and investment advice should have a certain degree of robustness and security, that is, to propose a safe and reliable investment strategy for users. Avoid risk-taking tendencies.

i. Financial Security and Compliance(FSC): The answer should conform to China's financial industry control system and China's market operation standards. For the financial industry and system interpretation standards that differ between China and other countries, the Chinese interpretation should be the only standard.

j. Risk Prediction and Control(RPC): The answer should make a clear judgment on the potential financial risks in the problem, and put forward the corresponding management and control strategies for the risks.

k. Financial Insight(FI): On the premise of making basic



Figure 4: An example of a problem in SC-Fin on professional analysis

judgments on the situation and development of the financial industry, the opinions and suggestions put forward by the answers need to have a certain height and depth, which can see the essence through the problems, and then assist users to make the best judgment.

Task type	Scoring standard
Fund Qualification	BR, IC, LI
Securities Qualification	BR, IC, LI
Banking Qualification	BR, IC, LI
Insurance Qualification (CICE)	BR, IC, LI
Economist	BR, IC, LI
Actuary	BR, IC, LI, MAE
Financial Planner	BR, LF, LI, SM
Futures Qualification	BR, IC, LI
Tax Accountant	BR, IC, LI
Certified Public Accountant (CPA)	BR, IC, LI
Financial Documents Summary	BR, LF, LI, FI
Financial Information Extraction	BR, IC
Financial Sentiment Judgment	BR, LI, CJ, FI
Mathematical Calculation	BR, IC, MAE
Logical Analysis	BR, LF, LI, CJ
Financial Compliance	BR, FSC, LI
Risk Management Application	BR, RPC, LI
Macroanalysis	BR, LF, LIy, CJ
Stock Analysis	BR, LF, LI, FI
Market Interpretation	BR, LF, LI, CJ
Financial Industry Analysis	BR, LF, LI, FI
Company Review	BR, LF, LI, AAB
Financial Report Comments	BR, LF, LI, AAB
Pan-investment Consultant Answer	BR, FSC, SM, FI
Fund Analysis	BR, FSC, SM, FI

Table 1: Correspondence between various tasks and corresponding evaluation criteria. BR: Basic Requirement, IC: Information Correctness, LF: Logic of Finance, LI: Language Intelligibility, MAE: Mathematical Ability and Efficiency, AAB: Ability to Analyze Business, CJ: Clear Judgment, SM: Safety Measure, FSC: Financial Security and Compliance, RPC: Risk Prediction and Control, FI: Financial Insight.

SuperCLUE-Fin

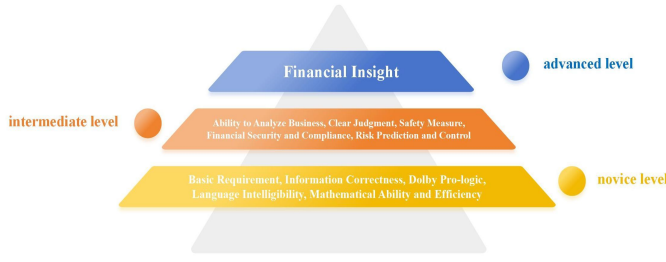


Figure 5: Overview of SuperCLUE-Fin evaluation criterion

Assessment and Scoring Methodology

Models are assessed through an API-based interaction, retrieving responses in either single or multi-turn conversational formats. For each question, the model’s answer is scored according to the defined criteria. After obtaining scores for all tasks, these are normalized to a percentage scale, representing the overall performance of the model.

Reliability Analysis

To verify the reliability and practicality of the SC-Fin benchmark, a pre-evaluation experiment was conducted on four representative models, randomly selecting 120 questions per model from a pool of over 1000 questions. Reviewers evaluated the model outputs using a quality classification system with levels of excellence, satisfactory, passable, and failing. The results showed a high level of consistency and reliability in the assessment process.

Model List	Reliability
Model One	0.9084
Model Two	0.9924
Model Three	0.8760
Model Four	0.9695
Average	0.9366

Table 2: Reliability Analysis

Grading Mechanism Explanation

SuperCLUE-Fin adopts a tiered grading mechanism to classify models into three distinct tiers based on their overall performance. Tier 1 models must achieve a minimum grade of A, with at least one of their basic or applied capability grades being A or above, denoting their competence in fulfilling both consumer (ToC) and business (ToB) requirements. Tier 2 models must have grades no lower than B in both basic and applied capabilities, indicating they meet ToC needs but may require enhancements for ToB purposes. Tier 3 models, graded C or below, indicate significant room for improvement in both areas.

By implementing such a rigorous and multifaceted evaluation system, the SuperCLUE-Fin benchmark not only serves as a platform to measure the progress and identify the shortcomings of existing Chinese-native financial LLMs

but also provides a roadmap for their future development, emphasizing the importance of improving database quality, refining the interpretation of financial problems based on domestic standards, and enhancing the models’ analytical and decision-making capacities in line with China’s financial context.

Score Range	Level	Tier
Above 75	A+	First Tier
70-75	A	First Tier
65-70	B	Second Tier
60-65	C	Third Tier
Below 60	D	Third Tier

Table 3: Grading Mechanism Explanation

Experimentation and Analysis

The experimentation phase of the SuperCLUE-Fin benchmark involved a systematic evaluation of various financial large language models across a multitude of tasks designed to test their understanding, reasoning, and applicability in the financial domain. The analysis delves into the models’ performances, strengths, weaknesses, and areas for improvement, providing valuable insights into the current state of domestic and international models in the context of Chinese financial services.

Model Name	Organization	Access
Baichuan2-13B-Chat	Baichuan	API
ChatGLM3-6B	ZhiPu	API
ERNIE Bot 4.0	Baidu	API
Gemma-7b-instruct	Google	API
GLM-4	ZhiPu	API
GPT-3.5 Turbo	OpenAI	API
GPT-4	OpenAI	API
GPT-4 Turbo	OpenAI	API
MoonShot-v1-128K	MoonShot	API
qwen-finance-14B	Alibaba	API
SparkDesk V3.5	Xunfei	API

Table 4: Model information

Performance Analysis

Model Name	Model level
GPT-4 Turbo	A+
MoonShot-v1-128K	A
GLM-4	A
SparkDesk V3.5	B
ERNIE Bot 4.0	B
GPT-4	B
GPT-3.5 Turbo	C
Baichuan2-13B-Chat	C
ChatGLM3-6B	D
qwen-finance-14B	D
Gemma-7b-instruct	D

Table 5: SuperCLUE-Fin Review - Overall Ranking. Note: The same level models are sorted alphabetically.

Upon assessment, it was observed that the overall financial capability of domestic models showed strong compet-

itiveness, with GLM-4 and MoonShot-v1-128k reaching the highest tier with an A-grade, though still trailing behind the leading GPT-4 Turbo. Meanwhile, models like iFlytek StarFire V3.5 and Wenxin Yiyuan 4.0 achieved a B-grade, demonstrating better performance compared to GPT-4. However, the majority of the models fell within the C-tier, suggesting there is substantial room for advancement in the maturity and functionality of domestic financial LLMs.

Model Name	Model level	Basic Capability	Applied Capability
GPT-4 Turbo	A+	A+	A+
GLM-4	A	A	B
MoonShot-v1-128K	A	A+	B
GPT-4	B	B	C
ERNIE Bot 4.0	B	B	B
SparkDesk V3.5	B	A	B
Baichuan2-13B-Chat	C	C	C
GPT-3.5 Turbo	C	B	C
ChatGLM3-6B	D	D	C
Gemma-7b-instruct	D	D	D
Average Level	B	B	B

Table 6: SuperCLUE-Fin Review - Summary of each ability level. Note: The same level models are sorted alphabetically.

The analysis further revealed disparities between the models' basic and applied capabilities. While they generally excelled in basic financial knowledge (such as financial theory, market regulations, and common financial instruments), there was a notable deficiency in their ability to apply this knowledge in more complex, real-world scenarios, requiring strategic decision-making, risk assessment, and compliance considerations.

Model Name	FF	FIEA	MCLA	FCRM	IR	IA
GPT-4 Turbo	A+	A+	A+	A+	A	A
GLM-4	A+	B	A	A	B	A
MoonShot-v1-128K	A+	B	B	A	B	B
SparkDesk V3.5	A	B	A	A	B	B
ERNIE Bot 4.0	A	B	C	A	B	B
GPT-4	A	B	B	B	C	B
GPT-3.5 Turbo	B	D	C	B	C	C
Baichuan2-13B-Chat	B	D	D	B	C	C
ChatGLM3-6B	C	D	D	B	C	C
qwen-finance-14B	C	C	D	C	D	C
Gemma-7b-instruct	D	C	D	C	D	D

Table 7: SuperCLUE-Fin Review - Summary of each task level (primary class). Note: The same level models are sorted alphabetically. FF: Financial Fundamentals, FIEA: Financial Information Extraction and Analysis, MCLA: Mathematical Calculation and Logical Analysis, FCRM: Financial Compliance and Risk Management, IR: Investment Research, IA: Investment Adviser.

In the specific task categories, models demonstrated varying degrees of success. For instance, in the financial understanding and cognition tasks, models could effectively summarize financial documents and extract pertinent information. However, in the realm of financial number-theoretic calculations and logical analysis, discrepancies became evident. The models' ability to handle complex financial computations and to provide detailed, contextually accurate explanations varied significantly, sometimes failing to account

for potential risks or limitations inherent in financial decisions, such as stock buybacks.

Taking the correlation Analysis between Financial Fundamentals and the tasks of Mathematical Calculation and Logical Analysis as an example, as shown in the figure below, the model shows a high correlation between these two tasks. Without considering the impact of the model's computing ability on the final result, the higher the model's mastery of financial knowledge is, the stronger the model's ability to solve the same financial mathematical calculation problems is. This shows that it is necessary to improve the quality of large financial model database and improve the interpretation standard of models for financial problems to improve the performance of large financial models.

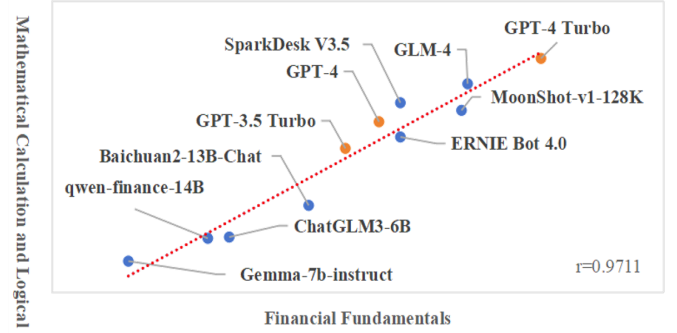


Figure 6: Financial Fundamentals - Mathematical Calculation and Logical Analysis

Performance analysis of each task

1) Financial Fundamentals

In the part of Financial Fundamentals, the grade of most models can reach B or above, which indicates that the models have a good grasp of basic knowledge in the financial field, and have the foundation to further expand the large model to deal with more complex financial problems and financial business.

Model Name	Model Level	FdQ	SQ	BQ	CICE	Ec	Ac	FP	FsQ	TA	CPA
GLM-4	A+	A+	A+	A+	A+	A+	B	A	A	A+	A+
GPT-4 Turbo	A+	A+	A+	A+	A+	A+	A+	A+	A+	A+	A+
MoonShot-v1-128K	A+	A+	A+	A+	A+	A+	B	A	A+	A+	A+
GPT-4	A	A	A	A	A	A	B	B	A+	A	B
ERNIE Bot 4.0	A	A	A+	A+	A+	A	C	A	A+	A	B
SparkDesk V3.5	A	A	A+	A+	A	A	B	A	A+	B	B
Baichuan2-13B-Chat	B	B	B	A+	B	A	D	B	C	A	B
GPT-3.5 Turbo	B	B	A	A+	A	B	D	B	A+	B	B
ChatGLM3-6B	C	C	C	A	B	D	D	C	C	C	C
qwen-finance-14B	C	C	C	C	B	D	D	C	D	D	C
Gemma-7b-instruct	D	D	D	C	D	D	D	D	D	D	D

Table 8: Financial Fundamentals Overview of model results. Note: The same level models are sorted alphabetically. FdQ: Fund Qualification, SQ: Securities Qualification, BQ: Banking Qualification, CICE: Insurance Qualification (CICE), Ec: Economist, Ac: Actuary, FP: Financial Planner, FsQ: Futures Qualification, TA: Tax Accountant, CPA: Certified Public Accountant (CPA)

2) Financial Information Extraction and Analysis In the part of Financial Information Extraction and Analysis, ex-

cept GPT-4 Turbo, the overall grade of other models is B or below, which indicates that there are still big problems in dealing with financial information in the financial large model. For example, it is not sensitive enough to capture financial related words, not enough to analyze financial professional words, and not enough to summarize and summarize financial terms skillfully.

Model Name	Model Level	FDS	FIE	FSJ
GPT-4 Turbo	A+	A	A+	A+
GLM-4	B	B	B	B
GPT-4	B	C	A+	B
MoonShot-v1-128K	B	B	B	B
ERNIE Bot 4.0	B	B	A	B
SparkDesk V3.5	B	B	B	B
Gemma-7b-instruct	C	D	C	C
qwen-finance-14B	C	C	D	D
Baichuan2-13B-Chat	D	C	D	C
ChatGLM3-6B	D	C	D	B
GPT-3.5 Turbo	D	C	D	C

Table 9: Financial Information Extraction and Analysis Overview of model results. Note: The same level models are sorted alphabetically. FDS: Financial Documents Summary, FIE: Financial Information Extraction, FSJ: Financial Sentiment Judgment

3) Mathematical Calculation and Logical Analysis In the part of Mathematical Calculation and Logical Analysis, all models show strong polarization phenomenon, that is, the model with higher comprehensive grade also shows better performance in mathematical calculation and logical analysis ability. However, the model with low comprehensive level showed significantly lower performance than the comprehensive ability in the above two types of tasks.

Model Name	Model Level	MC	LA
GPT-4 Turbo	A+	A+	A+
GLM-4	A	A	A
SparkDesk V3.5	A	B	A
GPT-4	B	B	B
MoonShot-v1-128K	B	B	A
GPT-3.5 Turbo	C	C	C
ERNIE Bot 4.0	C	D	A
Baichuan2-13B-Chat	D	D	B
ChatGLM3-6B	D	D	C
Gemma-7b-instruct	D	D	D
qwen-finance-14B	D	D	D

Table 10: Mathematical Calculation and Logical Analysis Overview of model results. Note: The same level models are sorted alphabetically. MC: Mathematical Calculation, LA: Logical Analysis

4) Financial Compliance and Risk Management In the part of Financial Compliance and Risk Management, each model almost shows no lower than the comprehensive level, which indicates that the financial model has a relatively mature ability in dealing with the problems related to laws and regulations and risk aversion and management. This is the key factor to ensure that the models do not break the law as they continue to evolve.

Model Name	Model Level	FC	RMA
GPT-4 Turbo	A+	A	A+
GLM-4	A	A	A
MoonShot-v1-128K	A	A	A
ERNIE Bot 4.0	A	A	A
SparkDesk V3.5	A	A	A
Baichuan2-13B-Chat	B	B	B
ChatGLM3-6B	B	B	B
GPT-3.5 Turbo	B	B	B
GPT-4	B	B	B
Gemma-7b-instruct	C	D	C
qwen-finance-14B	C	C	C

Table 11: Financial Compliance and Risk Management Overview of model results. Note: The same level models are sorted alphabetically. FC: Financial Compliance, RMA: Risk Management Application

5) Investment Research In the Investment Research section, the models almost all performed below their composite grade. Considering the specific content involved in this task type, the above phenomenon indicates that the ability of the financial model to deal with the market, the industry and the specific business within the company is not good, and it is expected that the task ability will be further improved with the further development of the model.

Model Name	Model Level	Ma	SA	MI	FIA	CR	FRC
GPT-4 Turbo	A	A	D	A	B	A	A
GLM-4	B	A	D	A	B	B	A
MoonShot-v1-128K	B	A	D	A	B	B	A
ERNIE Bot 4.0	B	A	D	B	D	B	A
SparkDesk V3.5	B	B	D	A	C	B	C
Baichuan2-13B-Chat	C	B	D	B	C	C	C
ChatGLM3-6B	C	B	D	C	D	D	D
GPT-3.5 Turbo	C	C	D	C	D	C	D
GPT-4	C	B	D	B	C	C	C
Gemma-7b-instruct	D	D	D	D	D	D	D
qwen-finance-14B	D	C	D	D	D	D	D

Table 12: Investment Research Overview of model results. Note: The same level models are sorted alphabetically. Ma: Macroanalysis, SA: Stock Analysis, MI: Market Interpretation, FIA: Financial Industry Analysis, CR: Company Review, FRC: Financial Report Comments

6) Investment Adviser In the Investment Adviser section, the performance of each model is similar to that in the Investment Research section, that is, the performance is not satisfactory when specific business QA is involved. In particular, when simulated investment advisers answer users' investment-related questions, they are not in-depth and detailed enough, and their suggestions are relatively vague.

Model Name	Model Level	PCA	FA
GLM-4	A	A	A+
GPT-4 Turbo	A	A	B
GPT-4	B	B	B
MoonShot-v1-128K	B	B	A
ERNIE Bot 4.0	B	B	A
SparkDesk V3.5	B	B	B
Baichuan2-13B-Chat	C	C	D
ChatGLM3-6B	C	C	C
GPT-3.5 Turbo	C	C	C
qwen-finance-14B	C	C	D
Gemma-7b-instruct	D	D	D

Table 13: Investment Adviser Overview of model results. Note: The same level models are sorted alphabetically. PCA: Pan-investment Consultant Answer, FA: Fund Analysis

Reliability and Validity

The reliability of the SuperCLUE-Fin benchmark was rigorously examined by conducting experiments on four representative models, with human evaluators reviewing a random subset of responses. The high rate of agreement between the model outputs and evaluator judgments confirmed the reliability of the assessment method. The scoring system was shown to be consistent and fair, with a clear distinction between model performance levels. From the empirical data and analysis, it can be concluded that although certain domestic models have reached commendable levels of competency, there is still a gap to bridge before attaining the same level of sophistication and adaptability as seen in top-performing international models like GPT-4 Turbo. Specifically, the specialized capabilities of financial LLMs in China need considerable enhancement, particularly in relation to application-oriented tasks and the provision of contextually precise, balanced, and risk-aware recommendations. It was also noted that the relationship between a model’s mastery of financial knowledge and its performance in financial calculations underscored the importance of having a robust financial knowledge base and adherence to consistent and accurate interpretations of financial principles, especially those aligned with Chinese financial regulations. Overall, the SuperCLUE-Fin benchmark not only serves as a powerful diagnostic tool to evaluate the strengths and weaknesses of financial LLMs but also provides a roadmap for future model development, emphasizing the necessity to upgrade databases, refine interpretations of financial problems, and bolster the models’ ability to navigate the intricacies of the Chinese financial landscape. This will ultimately lead to more effective, compliant, and customer-centric financial services powered by AI technologies.

3 Related Work

In the field of large language models (LLMs) and their application in the financial domain, several benchmarks and models have emerged globally. The FINQA dataset, is a large-scale collection of financial reports paired with expert-annotated questions and answers, aiming to enhance the automation of financial data analysis through complex numerical reasoning (Chen et al., 2021).

Another relevant contribution comes from the Financial Narrative Processing (FNP) shared task series, which has hosted annual competitions since 2018. The tasks involve extracting and summarizing financial information from earnings calls and annual reports, sentiment analysis, and event extraction (Chen et al., 2021). These efforts have significantly pushed the boundaries of natural language processing (NLP) techniques in the financial sector and provided valuable insights into the performance and limitations of contemporary NLP models in handling complex financial text.

On the model side, OpenAI’s GPT series, especially GPT-3.5, GPT-4 have demonstrated impressive results when fine-tuned for financial applications. They have been used for generating financial reports, giving investment advice, and assisting in compliance tasks (OpenAI et al., 2023). However, these models were primarily developed for a global English-speaking audience and might not fully capture the specificities and complexities of the Chinese financial market.

Our proposed SuperCLUE-Fin benchmark builds upon these prior works by tailoring the evaluation to the Chinese financial context. Unlike previous benchmarks, SC-Fin emphasizes native Chinese content and financial regulations, aiming to address the unique challenges faced by AI models operating within China’s financial ecosystem. It incorporates a diverse range of tasks, from fundamental financial knowledge to complex applications, ensuring that the tested models can not only comprehend and analyze financial data but also adhere to local regulatory guidelines and cultural nuances, thus paving the way for more accurate, reliable, and contextually relevant financial AI solutions in China.

4 Conclusion

The SC-Fin benchmark represents a significant step towards a more standardized, professional, and transparent assessment process for FLMs in the Chinese financial domain. Its rigorous testing methodology and scoring system not only provide a comprehensive measure of the current state of these models but also offer valuable insights for developers to improve upon. The introduction of a tiered grading mechanism further enhances the credibility of the assessment and aids stakeholders in gauging the suitability of different models for various use cases.

References

- [1] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.
- [2] Cobbe, Karl and Kosaraju, Vineet and Bavarian, Mohammad and Chen, Mark and Jun, Heewoo and Kaiser, Lukasz and Plappert, Matthias and Tworek, Jerry and Hilton, Jacob and Nakano, Reiichiro and others, 2021. Training verifiers to solve math word problems. arXiv:2110.14168.
- [3] Hendrycks, Dan and Burns, Collin and Kadavath, Saurav and Arora, Akul and Basart, Steven and Tang, Eric and Song, Dawn and Steinhardt, Jacob, 2021. Mea-

suring mathematical problem solving with the math dataset. arXiv:2103.03874.

- [4] Zheng, Lianmin and Chiang, Wei-Lin and Sheng, Ying and Zhuang, Siyuan and Wu, Zhanghao and Zhuang, Yonghao and Lin, Zi and Li, Zhuohan and Li, Dacheng and Xing, Eric and others, 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.
- [5] Li, Xuechen and Zhang, Tianyi and Dubois, Yann and Taori, Rohan and Gulrajani, Ishaan and Guestrin, Carlos and Liang, Percy and Hashimoto, Tatsunori B, 2023. AlpacaEval: An automatic evaluator of instruction-following models. GitHub repository.
- [6] Keisuke Sakaguchi and Ronan Le Bras and Chandra Bhagavatula and Yejin Choi, 2019. An Adversarial Winograd Schema Challenge at Scale. arXiv:1907.10641.
- [7] Peter Clark and Isaac Cowhey and Oren Etzioni and Tushar Khot and Ashish Sabharwal and Carissa Schoenick and Oyvind Tafjord, 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

A Examples for problem in SC-Fin

Below are examples of problems in SuperCLUE-Fin.

Basic Capability

Example 1: Fund Qualification

Question: The net value of QDII fund issues regulation: QDII fund has any special requirements in net value issues?

Follow-up: What is the impact of these regulations on investor transparency and confidence?

Example 2: Actuary

Question: An insurance company uses claims data to predict future health insurance costs. If the average claim rate over the past five years is 0.025, the average claim amount is 1,500 dollars, and the claim rate is expected to rise by 0.1 in the future, how should the company adjust its premium?

Follow-up: If claims are expected to rise by 0.2, how much do premiums need to adjust to account for higher costs?

Example 3: Financial Planner

Question: How to optimize the portfolio to enhance its resilience to risk?

Follow-up: Please explain in detail the difference between asset allocation and diversification and their importance.

Example 4: Financial Documents Summary

Question: Please answer the question based on the information you already know. The following is a summary summary. Requirements: no more than 150 words, focus, organized content.

Known information: Amperon (301413) is mainly engaged in temperature and pressure sensors, and the downstream is mainly automotive electronics and home appliances. The company's product line includes thermistors and temperature sensors, oxygen sensors, pressure sensors, downstream applications are mainly in automotive, home appliances, industrial energy storage, etc. In recent years, the proportion of pressure sensor revenue has increased significantly.

... ; middle omission ;...

In addition, the company has obtained a number of MEMS pressure sensor projects from a well-known European OEMs customer, some of which have realized SOP in the first quarter of 2024.

Example 5: Mathematical Calculation

Question: Portfolio diversification: Xiaohong invests in four different stocks, and the investment proportion and expected return rate of each stock are 0.3 and 0.05 for A stock, 0.3 and 0.06 for B stock, 0.2 and 0.07 for C stock, and 0.2 and 0.08 for D stock respectively. The expected return rate of her portfolio is calculated.

Follow-up: If the expected return on A-stock falls to 0.03, how will the expected return on the entire portfolio adjust?

Example 6: Logical Analysis

Question: The ROE of a listed company is higher than 0.25 for three consecutive years, but its P/E ratio is lower than the industry average. The company's management believed that the market was undervaluing the company and decided to initiate a share buyback program. Q. What is the logic behind the company's move? Can share repurchases effectively boost stock prices?

Follow-up: What other factors should be considered in this decision?

Applied Capabilities

Example 7: Financial Compliance

Question: What are the regulatory policies for high-frequency trading in the securities market?

Follow-up: How do these policies affect market fairness and transparency?

Example 8: Financial Industry Analysis

Question: Answer the question based on known information. Known information: Title: Coal mining industry tracking weekly report: off-season demand continues to be weak coal prices stabilize and shock. Source: Soochow Securities.

Content: Industry status: thermal coal, off-season demand remains weak, port coal prices stabilize and shock. This week (April 7 to April 12), the spot price of thermal coal at the port fell by 15 yuan/ton month-on-month to 801 yuan/ton. ... ; middle omission ;... We maintain the "overweight" rating of the industry, recommend high flexibility targets China Grand Energy and China Grand Logistics, and suggest paying attention to Haohua Energy; In addition, we still recommend insurance OCI capital allocation of high dividend targets: China Shenhua, Shaanxi Coal Industry, it is

suggested to pay attention to Yankuang Energy. Risk warning: downstream demand is less than expected; Coal prices fell sharply as supply protection was stronger than expected. Excuse me, based on the current market trend, why the continued optimism for coal prices in the first half of 2024? Follow-up: What is the current supply and demand situation of the coal market and how does it affect the trend of coal prices?

Example 9: Company Review

Question: Answer the question based on known information. The following is the content of the article: On April 11, Yum China (HK 09987, HK 299.8, market value HK 117.9 billion) released an open letter to shareholders, in which CEO Qu Cuirong mentioned that the company plans to accelerate the pace of returning at least US 3 billion to shareholders in the next three years ... j middle omission i... At the same time, our flexible store model helps us reduce our upfront investment costs, and our strong own supply chain management capabilities and logistics network help us reach remote areas, all of which give us a strong competitive advantage. Please, Yum China's operating profit in 2023 will reach 1.1 billion US dollars, with core operating profit growing 0.79. Based on the company's store expansion plan and lower-tier city layout in the next three years, how do you think its profitability and valuation are reasonable? What are the advantages over other companies in the same industry? Follow-up: What do the frequent debt problems of Beingmate Group reflect the deficiencies in corporate governance and fund management? What improvement measures does the company need in financing and investment decisions in the future?

Example 10: Pan-investment Consultant Answer

Question: Futures investment: recently the price of bulk commodities fluctuates greatly, how should individual investors participate in the futures market?

Follow-up: Compared with securities investment, what are the characteristics of futures investment in terms of leverage ratio and trading mechanism? What are the risks and benefits?