BizFinBench.v2:
Towards Reliable LLMs in Finance via Real-User Data and Offline/Online Bilingual Evaluation

Xin Guo[1,2,*], Rongjunchen Zhang[1,*,♠], Guilong Lu[1], Xuntao Guo[1],
Shuai Jia[1], Zhi Yang[2], Liwen Zhang[2,♠]

[1] HiThink Research
[2] Shanghai University of Finance and Economics

CONTENT

01

Introduction

Introduction

Limitations of Existing Financial Benchmarks

The core characteristics of financial scenarios **lie in authenticity and online capability**, yet the vast majority of current financial benchmarks are undermined by two fundamental limitations:

- **Detached from Real Business Scenarios** (e.g., FinMaster, CFLUE) Most existing benchmarks rely on simulated data or generic datasets, lacking the complex logic and context of real-world financial operations, leading to a severe disconnect between evaluation results and practical deployment performance.
- **Focusing Solely on Static Offline Tasks** (e.g., FinBen, FinanceReasoning) Evaluation tasks are primarily static Q&A or analysis, neglecting critical online capabilities such as real-time market analysis, dynamic risk control, and portfolio optimization.

BizFinBench.v2: Our Solution

By simulating real-world document perturbations and multi-turn dialogues, BizFinBench.v2 captures the full complexity of financial workflows. The work's core contributions include:

- **Real Business Data-Driven:** Constructed from authentic user query-response data from Chinese and US stock markets, ensuring the authenticity and practicality of evaluation scenarios.
- **Offline + Online Dual-Track Evaluation:** Innovatively integrates 8 offline core capability tasks and 2 online real-time performance tasks for a comprehensive assessment of model strength.
- **Expert-Level Error Analysis:** Beyond performance metrics, it deeply dissects the causes of model failures to provide clear directions for optimization.

02

Related Work

Related Work

Financial Business Analysis(Simulation → Real)

Gap 1 – Simulated Data Sources

- Current financial benchmarks (e.g., FinEval, CFLUE, FinQA) rely heavily on **simulated/generic text data**, lacking **authentic user queries** from real Chinese/U.S. equity markets. They ignore real business logic and workflow complexity, leading to a **huge gap between benchmark performance and real-world efficacy.**

Gap 2 – Static Offline-Only Evaluation

- Existing benchmarks **focus exclusively on static offline tasks** (e.g., financial QA, report analysis), with **no support for online real-time tasks** (e.g., stock prediction, portfolio allocation). They fail to assess the online capability critical to real financial services.

Gap 3 – Monolingual & Narrow Scenario Coverage

- Most benchmarks **are monolingual** (Chinese/English only) and cover **limited financial scenarios**, **lacking cross-lingual evaluation and full-cycle business coverage.** They cannot reflect real cross-border financial business demands.

BizFinBench.v2

To fix above gaps, BizFinBench.v2 is designed as:

- **Real Business Data** (RBD)
- **Online Testing**(OT)
- **Cross-Lingual**(CL)

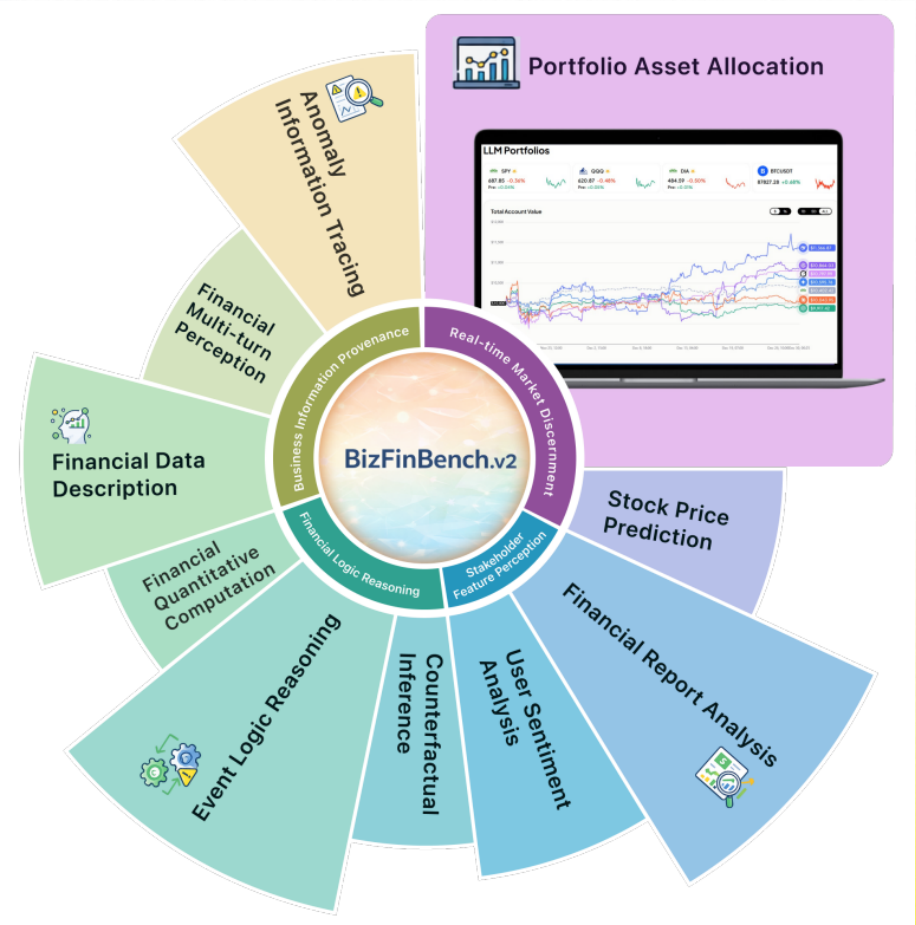
Benchmarks	CL	OT	RBD
FinEval (Guo et al., 2025b)	✗	✗	✗
CFLUE (Zhu et al., 2024)	✓	✗	✗
FinQA (Chen et al., 2021)	✗	✗	✗
ConvFinQA (Chen et al., 2022)	✗	✗	✗
FinMaster (Jiang et al., 2025)	✗	✗	✗
FinBen (Xie et al., 2024)	✓	✗	✗
FinanceMath (Zhao et al., 2024)	✗	✗	✗
FinanceReasoning (Tang et al., 2025)	✗	✗	✗
BizFinBench.v2	✓	✓	✓

03

BizFinBench.v2 Benchmark

3.1 BizFinBench.v2 Framework

- **Business Information Provenance：** Focuses on real user-facing financial inquiries, requiring LLMs to identify useful information, remove noise, and verify financial facts from heterogeneous sources.
- **Financial Logic Reasoning：** Evaluates whether LLMs can conduct financial computation, event reasoning, and counterfactual analysis based on market data and financial logic.
- **Stakeholder Feature Perception：** Tests whether LLMs can understand users, companies, industries, and market participants through sentiment analysis and financial report analysis.
- **Real-time Market Discernment：** Extends evaluation from static offline tasks to online market environments, including stock price prediction and portfolio asset allocation.



3.2 BizFinBench.v2 Introduction

Scenarios	Tasks	Avg. Input Tokens	#Questions
Business Information Provenance	Anomaly Information Tracing	8679	3963
	Financial Multi-turn Perception	10361	4497
	Financial Data Description	3577	3803
Financial Logic Reasoning	Financial Quantitative Computation	1984	2000
	Event Logic Reasoning	437	3944
	Counterfactual Inference	2267	604
Stakeholder Feature Perception	User Sentiment Analysis	3326	4000
	Financial Report Analysis	19681	2000
Real-time Market Discernment	Stock Price Prediction	5510	4049
	Portfolio Asset Allocation	–	–
All	–	–	**28860**

- BizFinBench.v2 includes **8 offline tasks** and **2 online tasks**, comprising a total of **28,860 questions**.

- BIP comprises 12,263 questions, including **3,963 AIT tasks**, **4,497 FMP tasks**, and **3,803 FDD tasks**. It evaluates LLMs' ability to filter key clues from massive multi-dimensional information, exclude interference, integrate historical user queries, and judge the authenticity and accuracy of financial data.
- FLR consists of 6,548 questions, containing **2,000 FQC tasks**, **3,944 ELR tasks**, and **604 CI tasks**. It evaluates rigorous financial reasoning, accurate indicator calculation, chronological and causal event reasoning, and counterfactual inference based on user-proposed hypotheses.
- SFP includes 6,000 questions, divided into **4,000 SA tasks** and **2,000 FRA tasks**. It aims to provide users with in-depth analysis and summaries of the market or industry, supporting customized services, highly relevant product recommendations, and enterprise industry ranking analysis.
- RMD covers two online tasks: Stock Price Prediction and Portfolio Asset Allocation. Since SPP utilizes the same accuracy metric as offline tasks, have **4,049 questions** generated during the evaluation process.

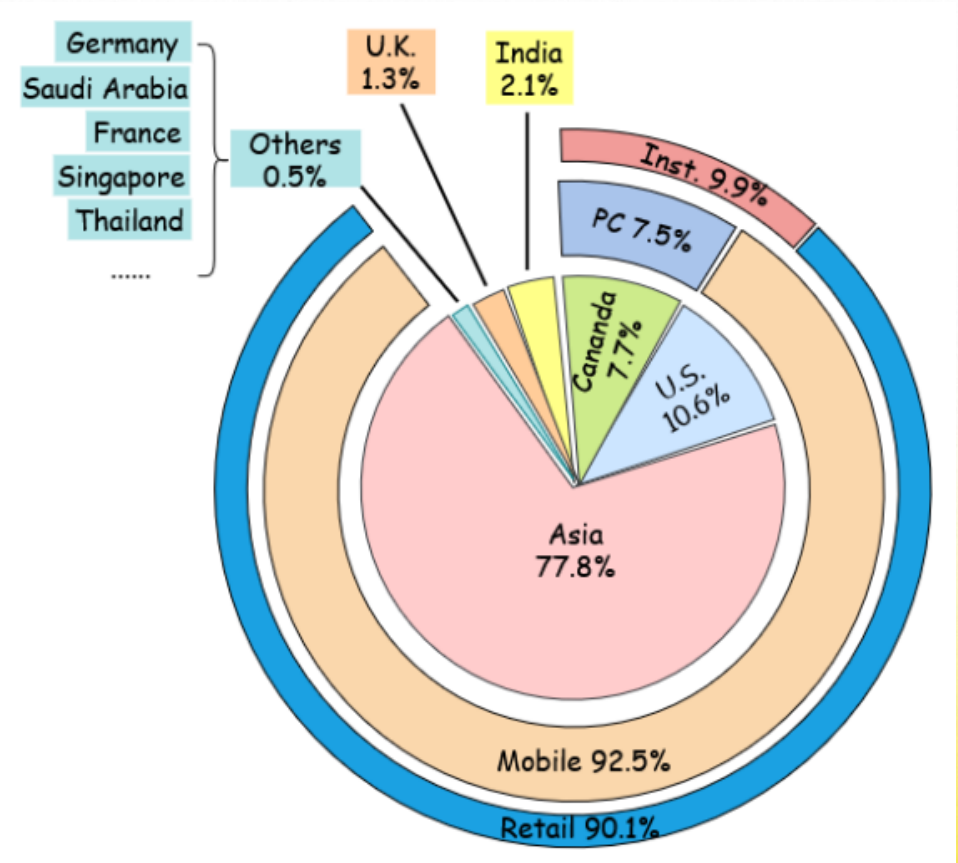
3.2 BizFinBench.v2 Introduction

Data Construction

- BizFinBench.v2 is built upon authentic user query-response data from **Chinese and U.S. equity markets**. All data is derived from real financial business platforms and authentic user demands within financial business scenarios.
- All questions in BizFinBench.v2 are verifiable **open-ended questions**, rather than simple multiple-choice or true/false questions.

Quality Control

- Offline Tasks: **Three-Level Progressive Mechanism**
  - Stage 1: Platform Clustering and Desensitization
  - Stage 2: Frontline Staff Review
  - Stage 3: Expert Team Cross-Validation
- Online Tasks: **Expert-Defined Configuration**
  - the structured data required for the Stock Price Prediction task
  - the specific equity market configurations for the Portfolio Asset Allocation task
  - the system prompts



04

Experiments Settings

Experiments Settings

We **evaluated 21 large language models** , with close-source models accessed through their respective APIs and open-source models deployed locally. All inference tasks were run on 8×NVIDIA H200 GPUs. We additionally invited **two financial experts** who were not involved in the data construction process to participate in the competition, so as to **conduct a comparative analysis with the performance of LLMs.**

Category	Model	Creator	Parameter	Access	Version Date	Domain
Proprietary	GPT-5	OpenAI	Undisclosed	API	2025.11	General
	Gemini-3	Google	Undisclosed	API	2025.11	General
	Kimi-k2	MoonshotAI	Undisclosed	API	2025.11	General
	Claude-Sonnet-4	Anthropic	Undisclosed	API	2025.9	General
	Doubao-Seed-1.6	ByteDance	Undisclosed	API	2025.6	General
	Grok-4	X AI	Undisclosed	API	2025.7	General
	Qwen3-Max	Alibaba Cloud	Undisclosed	API	2025.4	General
Open-Source	Qwen2.5-7B-Instruct	Alibaba Cloud	7B	Weights	2024.9	General
	Qwen2.5-72B-Instruct	Alibaba Cloud	72B	Weights	2024.9	General
	Qwen3-32B	Alibaba Cloud	32B	Weights	2025.4	General
	Qwen3-235B-A22B-Thinking-2507	Alibaba Cloud	235B	Weights	2025.4	General
	InternLM2.5-7B	Shanghai AI Laboratory	7B	Weights	2025.3	General
	InternLM2.5-20B	Shanghai AI Laboratory	20B	Weights	2025.3	General
	GLM-Z1-9B	ZhipuAI	9B	Weights	2025.4	General
	GLM-Z1-32B	ZhipuAI	32B	Weights	2025.4	General
	DeepSeek-R1-Distill-Qwen-7B	DeepSeek AI	7B	Weights	2025.2	General
	DeepSeek-R1-Distill-Qwen-32B	DeepSeek AI	32B	Weights	2025.2	General
	DeepSeek-R1	DeepSeek AI	671B	Weights	2025.12	General
	Fin-R1	Shanghai University of Finance and Economics	7B	Weights	2025.3	Financial
	FinX1	Duxiaoman-DI	70B	Weights	2024.12	Financial
	Dianjin-R1	Alibaba Cloud	32B	Weights	2025.4	Financial
	Fino1	The Fin AI	14B	Weights	2025.2	Financial

Evaluation Method

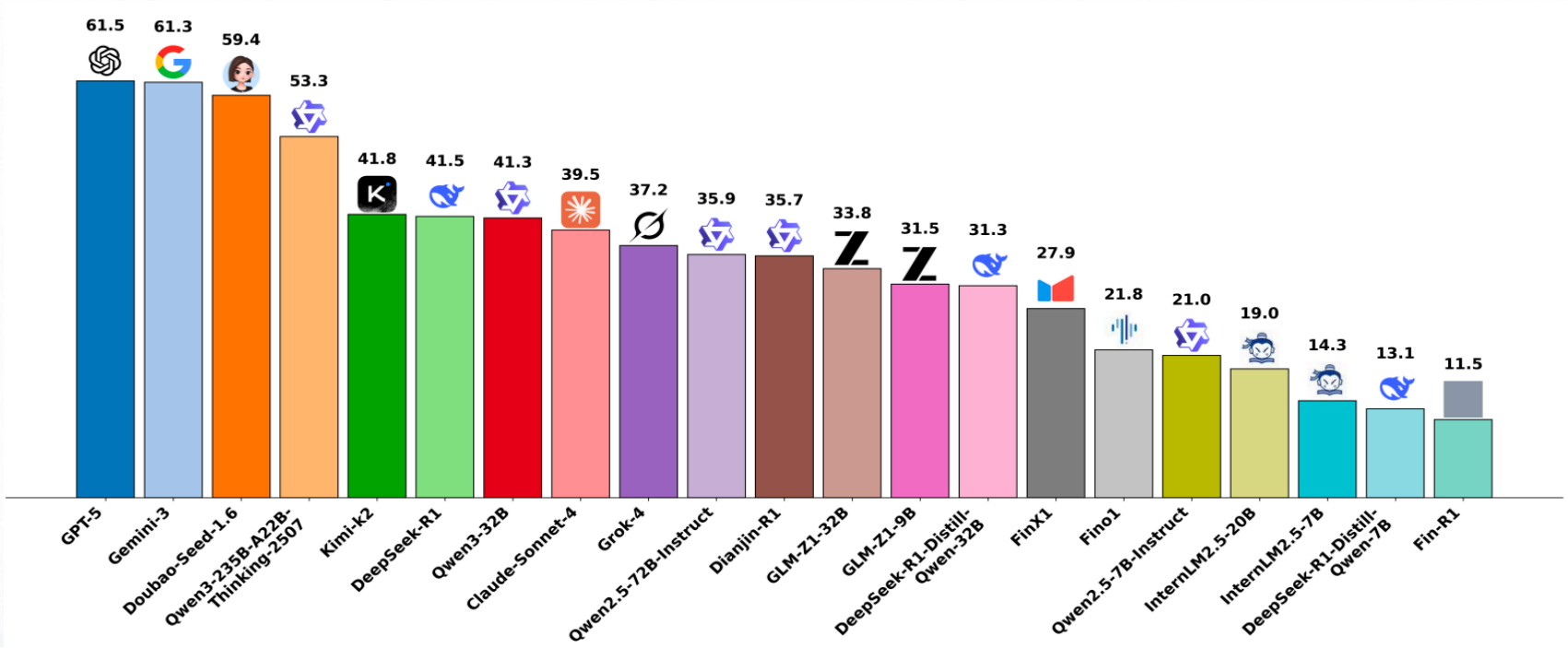
(1) Regarding SA and SPP tasks, we require the LLM to output a **prediction interval** based on its predicted value, consistent with the error tolerance of actual business operations. Correctness is determined by verifying whether the provided prediction interval contains the ground truth. Specifically, the business tolerance for the SA task is set at 10%, while the tolerance for the SPP task is set at 1%.

(2) The evaluation metrics for the PAA task primarily focus on investment indicators such as cumulative return, Sharpe ratio, and maximum drawdown. Therefore, the models evaluated for this task (mainly the six currently most powerful proprietary commercial models) and their corresponding experimental results are independent of other tasks.

05

Results

5.1 Experiment Results



- BizFinBench.v2 conducted zero-shot evaluations on **21** mainstream large language models . **GPT-5** ranked first with an overall accuracy of 61.5% .
- Gemini-3 and Doubao-Seed-1.6 performed outstandingly, ranking among the top three in multiple tasks. Among open-source models, Qwen3-235B-A22B-Thinking-2507 performed the best, reaching an average accuracy of 53.3%. In contrast, the highest average accuracy among financial domain models was only 35.7%, which is 5.6% lower than Qwen3-32B of the same parameter scale.
- Two factors: their **training data is centered on open-source financial datasets**, making it difficult to cover the complex characteristics of real-world financial business. their **business coverage is narrow** (e.g., only incorporating customer service Q&A data), rendering them unable to adapt to actual scenarios with high volatility and long contexts.

5.1 Experiment Results

Model	Size	AIT	FMP	FDD	FQC	ELR	CI	SA	FRA	SPP	Average
Proprietary LLMs											
GPT-5	unknown	54.2	90.8	68.3	89.2	62.0	83.9	18.8	54.1	32.1	61.5
Gemini-3	unknown	64.8	87.0	69.7	85.8	69.5	82.2	7.4	50.8	34.9	61.3
Doubao-Seed-1.6	unknown	62.6	90.2	63.8	78.2	61.2	78.7	22.7	46.3	31.1	59.4
Kimi-k2	unknown	55.2	80.9	22.4	62.2	44.6	15.4	20.1	45.0	30.5	41.8
Claude-Sonnet-4	unknown	54.8	79.9	28.4	29.8	44.8	23.4	17.3	47.7	29.1	39.5
Grok-4	unknown	61.8	86.6	37.4	9.3	42.1	4.7	17.8	45.2	30.3	37.2
Open-source General LLMs											
Qwen3-235B-A22B-Thinking-2507	235B	49.3	87.9	68.0	76.0	50.6	72.2	16.4	22.5	36.9	53.3
DeepSeek-R1	671B	58.9	87.2	42.1	21.7	48.9	8.1	23.9	50.0	32.3	41.5
Qwen3-32B	32B	54.3	80.9	40.0	48.4	42.0	47.2	13.4	40.5	5.1	41.3
Qwen2.5-72B-Instruct	72B	61.0	78.2	26.5	19.8	39.9	20.7	21.6	47.0	8.2	35.9
GLM-Z1-32B	32B	49.6	66.9	45.6	34.4	31.6	31.3	1.2	36.8	6.8	33.8
GLM-Z1-9B	9B	46.5	69.2	40.4	26.9	35.2	25.2	0.4	36.7	3.4	31.5
DeepSeek-R1-Distill-Qwen-32B	32B	52.1	75.4	20.5	21.0	35.9	24.2	18.4	27.3	6.8	31.3
Qwen2.5-7B-Instruct	7B	35.2	43.2	33.8	0.8	20.9	1.0	23.3	28.3	2.8	21.0
InternLM2.5-20B	20B	32.1	41.1	32.0	0.2	27.8	0.7	3.7	32.5	0.6	19.0
InternLM2.5-7B	7B	28.3	14.8	30.6	0.5	18.9	0.3	6.7	28.5	0.0	14.3
DeepSeek-R1-Distill-Qwen-7B	7B	16.6	30.9	17.1	3.5	10.3	6.2	19.1	13.0	1.0	13.1
Open-source Financial LLMs											
Dianjin-R1	32B	54.2	70.7	40.9	25.3	45.8	22.0	6.7	46.0	9.9	35.7
FinX1	70B	47.9	73.0	29.5	14.3	31.5	11.6	5.3	34.7	3.6	27.9
Fino1	14B	27.5	38.6	24.3	14.1	14.6	22.1	11.1	35.3	8.6	21.8
Fin-R1	7B	21.2	29.1	0.5	1.5	9.8	3.2	10.9	24.5	2.9	11.5
Financial Experts	–	92.6	98.0	94.5	100	91.7	100	57.9	96.0	32.3	84.8

Model	TR	PF	SR	MD	TA
LLM					
DeepSeek-R1	+47.34%	1.80	1.25	-53%	14734
Claude-Sonnet-4	-2.11%	0.90	-0.05	-11%	9789
Qwen3-Max	-3.79%	0.85	-0.10	-13%	9621
Gemini-3	-6.61%	0.83	-0.20	-13%	9339
Grok-4	-10.15%	0.76	-0.32	-15%	8985
GPT-5	-13.80%	0.72	-0.45	-16%	8620
Quantitative Strategy					
MA5-MA20	-1.61%	1.13	-3.97	-2.21%	9839
EW_week	22.41%	4.74	2.21	-3.05%	12241
EW_month	22.08%	3.34	1.68	-6.39%	12208
EW_quarter	32.98%	1.96	2.19	-9.20%	13298
M_Top5	101.71%	2.38	1.60	-56.89%	20171
M_Top10	78.85%	1.39	1.65	-41.51%	17885
Baseline					
SPY	–	–	–	–	10632

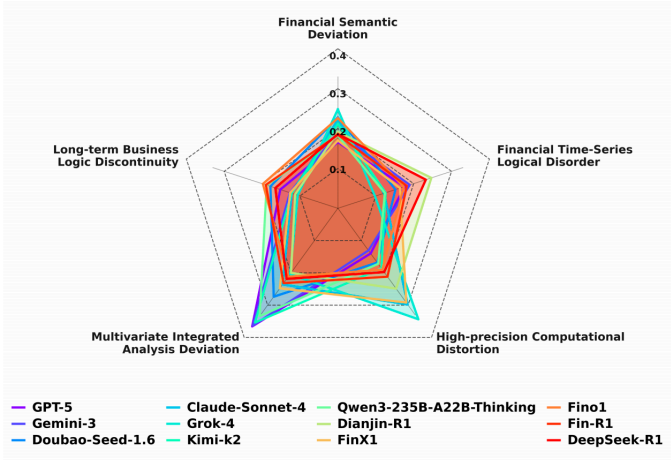
- The performance of LLMs in real-world financial business scenarios **remains significantly below the standards required for practical application (84.8%)**. This indicates that their performance in specialized financial tasks warrants critical attention. Conversely, the robust performance of DeepSeek-R1 in the PAA task underscores the latent potential of LLMs within the domain of commercial investment.
- **Models with larger parameters performed better in high-precision demand tasks** such as FDD, FQC, and CI. But Claude-Sonnet-4 and Grok-4 performed poorly in these tasks, indicating that their computational capabilities still need optimization to adapt to financial scenarios.
- The SA task exposed the models' shortcomings in relatively subjective analysis, even the best-performing DeepSeek-R1 achieved an average accuracy of only 32.3%. Furthermore, in the SPP task, the top model Qwen3-235B-A22B-2507 reached an accuracy of only 36.9%.

5.1 Experiment Results

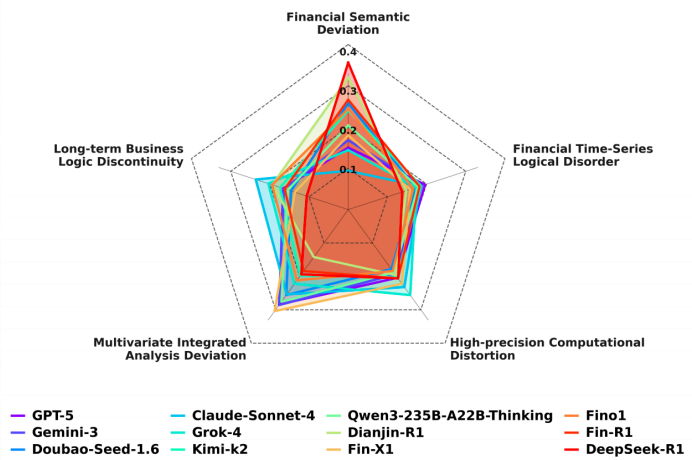
Model	Size	AIT	FMP	FDD	FQC	ELR	CI	SA	FRA	SPP	Average
Proprietary LLMs											
Doubao-Seed-1.6	unknown	64.3	88.1	61.2	79.2	60.0	75.7	20.3	47.8	6.7	56.9
Gemini-3	unknown	67.8	91.8	54.9	84.6	67.9	80.6	3.7	51.0	0.9	56.5
GPT-5	unknown	54.1	88.6	49.3	86.7	59.5	83.6	8.8	55.7	6.0	54.6
Kimi-k2	unknown	63.7	81.4	27.2	40.8	45.0	42.4	17.8	48.3	2.6	40.1
Grok-4	unknown	63.2	88.7	39.2	20.5	43.5	20.2	18.3	44.7	7.7	37.7
Claude-Sonnet-4	unknown	21.8	39.3	13.4	9.0	13.1	8.1	2.9	48.3	1.6	13.7
Open-source General LLMs											
Qwen3-235B-A22B-Thinking-2507	235B	48.4	88.5	63.6	75.3	58.4	70.6	12.8	21.7	4.2	52.7
DeepSeek-R1	671B	61.7	87.2	43.2	54.2	50.9	55.9	21.6	46.5	32.8	50.4
Qwen3-32B	32B	55.1	79.7	12.9	44.7	41.4	43.9	9.3	39.8	3.8	36.4
Qwen2.5-72B-Instruct	72B	60.9	82.0	31.0	28.8	39.6	27.2	13.0	45.2	7.6	36.3
DeepSeek-R1-Distill-Qwen-32B	32B	54.4	77.3	22.1	19.7	40.6	20.2	17.8	27.3	6.7	32.4
GLM-Z1-32B	32B	50.9	62.4	42.8	32.7	28.8	30.6	0.8	35.7	6.8	32.0
GLM-Z1-9B	9B	45.9	72.9	41.3	24.8	32.7	25.2	0.6	36.2	3.4	30.9
Qwen2.5-7B-Instruct	7B	26.1	49.4	31.9	4.9	21.1	5.0	16.3	28.5	2.5	19.6
InternLM2.5-20B	20B	33.0	58.1	17.4	1.7	29.4	2.0	3.0	33.7	0.4	18.1
DeepSeek-R1-Distill-Qwen-7B	7B	14.4	30.3	16.0	3.0	10.6	3.8	6.5	13.0	0.9	10.7
InternLM2.5-7B	7B	26.0	23.8	4.8	1.7	19.3	0.7	5.8	27.0	0.1	10.2
Open-source Financial LLMs											
Dianjin-R1	32B	53.0	76.6	39.7	18.0	41.3	20.0	5.6	30.0	5.8	32.5
FinX1	70B	45.8	74.4	25.3	8.5	32.0	7.1	2.5	34.7	3.0	24.8
Fino1	14B	36.5	58.4	0.0	13.5	28.2	8.9	12.8	28.5	8.8	20.9
Fin-R1	7B	22.7	30.8	35.7	1.8	10.7	3.6	7.5	24.2	0.6	14.2

- CoT amplified the reasoning logic defects in most models while releasing the potential of a few. Contrary to the zero-shot setting, the CoT setting **did not improve model performance but instead led to a decline for most models**, with some experiencing significant decay.
- Claude-Sonnet-4's average accuracy plummeted from 39.5% to 13.7%, highlighting its weakness in reasoning tasks. Conversely, DeepSeek-R1 achieved buck-the trend growth, with its average accuracy increasing by nearly 9% compared to zero-shot, reflecting better adaptability to CoT tasks.

5.2 Error Analysis



(a) Zero-Shot Error Distribution



(b) Chain-of-Thought Error Distribution

- A systematic dissection of the models' erroneous responses across multiple dimensions, including computational capability, semantic understanding, business logic, information integration, and temporal cognition.
- **Financial Semantic Deviation**: the model fails to accurately grasp the specific implications and impacts of key terms, numerical relationships, or dynamic changes within actual business scenarios.
- **Long-term Business Logic Discontinuity**: the models struggle to maintain a complete, coherent, and business rule-compliant logical chain when handling complex business analyses.
- **Multivariate Integrated Analysis Deviation**: the model struggles to effectively identify, weigh, and integrate the complex correlations and weak signals among different information sources.
- **High precision Computational Distortion**: the model fails to stably and reliably perform complex operations or quantitative deductions.
- **Financial Time-Series Logical Disorder**: the model fails to accurately identify and follow the critical chronological order and causal correlations embedded within them.

Thank you for your listening.