Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Correspondence:

When DeepSeek-R1 meets financial applications: benchmarking, opportunities, and limitations

Shuoling LIU[†1,2], Liyuan CHEN[†2], Jiangpeng YAN[†2,4], Yuhang JIANG[2],
Xiaoyu WANG[2], Xiu LI[†4], Qiang YANG[†‡1,3]

[1]*The Hong Kong University of Science and Technology, Hong Kong 999077, China*
[2]*E Fund Management Co., Ltd., Guangzhou 510000, China*
[3]*Webank AI, Shenzhen 518054, China*
[4]*Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China*
[†]E-mail: liushuoling@efunds.com.cn; chenly@efunds.com.cn; yanjiangpeng@efunds.com.cn; li.xiu@sz.tsinghua.edu.cn; qyang@cse.ust.hk

Received Apr. 10, 2025; Revision accepted Apr. 28, 2025; Crosschecked June 4, 2025; Published online June 27, 2025

https://doi.org/10.1631/FITEE.2500227

How the recent progress of reasoning large language models (LLMs), especially the new open-source model DeepSeek-R1, can benefit financial services is an underexplored problem. While LLMs have ignited numerous applications within the financial sector, including financial news analysis and general customer interactions, DeepSeek-R1 further unlocks the advanced reasoning ability with multiple reinforcement learning-integrated training steps for more complex financial queries and provides distilled student models for resource-constrained scenarios. In this paper, we first introduce the technological preliminaries of DeepSeek-R1. Subsequently, we benchmark the performance of DeepSeek-R1 and its distilled students on two public financial question–answer (QA) datasets as a starting point for interdisciplinary research on financial artificial intelligence (AI). Then, we discuss the opportunities that DeepSeek-R1 offers to current financial services, its current limitations, and three future research directions. In conclusion, we argue for a proper approach to adopting reasoning LLMs for financial AI.

‡ Corresponding author
ORCID: Shuoling LIU, https://orcid.org/0009-0003-1960-3004; Qiang YANG, https://orcid.org/0000-0001-5059-8360

1 Introduction

Fueled by Web-scale datasets, high-performance computing devices, and deep learning algorithms, recent years have witnessed the emergence of LLMs (Chang et al., 2024), which are paving the way toward artificial general intelligence (AGI) (Zhou et al., 2024). Built on the Transformer (Vaswani et al., 2017) architecture and trained with extensive text data corpora, GPT-3 (Brown et al., 2020), released by OpenAI in 2020, first demonstrated an unprecedented ability to understand, generate, and manipulate human language in a manner that closely mimics human-like intelligence, with a number of follower LLMs including Google's Gemini (Gemini Team of Google, 2023) and Alibaba's Qwen (Bai JZ et al., 2023). However, early LLMs struggled with hallucination problems (Huang et al., 2025) and usually failed to solve complex reasoning problems. With the notable breakthrough of the chain-of-thought (CoT) prompting technique (Wei et al., 2022), OpenAI released the first reasoning LLM called OpenAI-o1 (Jaech et al., 2024) in 2024 by training this model to increase the length of self-generated step-by-step human-like reasoning processes during the post-training process. Despite OpenAI-o1's remarkable capabilities, which can achieve human-level or superior performance on various tasks (including competitive programming, advanced mathematics, and scientific research), access to OpenAI-o1 is exclusively available to a limited number of subscribers, restricting its widespread utilization and impact across broader communities. In response to this demand, the introduction of DeepSeek-R1 (Guo DY et al., 2025) in January 2025 marked a significant milestone. As the first open-weight reasoning LLM that can achieve a similar performance as OpenAI-o1, DeepSeek-R1 gained global attention by democratizing access to advanced AI technologies, making them nearly cost-free. To take it a step further, DeepSeek also open-sourced a series of distilled student models that can be deployed in resource-constrained conditions.

DeepSeek-R1 and its deployment-friendly distilled student models have fostered swift applications in various fields (Peng et al., 2025; Qian et al., 2025), where financial contexts arguably demand the highest levels of reasoning capabilities and data-processing prowess due to their inherent complexity and risk. In the FinTech sector, while models such as BloombergGPT (Wu et al., 2023) and XuanYuan (Zhang XY and Yang, 2023) have been previously developed specifically for finance, they often fall short when it comes to sophisticated reasoning tasks such as finance numeric analysis and law and regulation logic thinking. This paper aims to discuss DeepSeek-R1 from the perspective of the FinTech industry. Given its open-source nature and advanced reasoning abilities, DeepSeek-R1 offers a promising foundation upon which more specialized models tailored to the exacting demands of financial service can be constructed.

In this paper, we first introduce the technological preliminaries of DeepSeek-R1, emphasizing its contributions to the multistep reinforcement learning (RL)-based training strategy and the exploration of distilling reasoning ability using student models with fewer parameters. Subsequently, we benchmark the performance of DeepSeek-R1 and its distilled students on two public financial QA datasets as a starting point for interdisciplinary research on financial AI, compared to common baseline non-reasoning LLMs. Then, we discuss how DeepSeek-R1 can benefit current financial services with its advanced reasoning ability. While acknowledging the strides made, we conclude that there are potential drawbacks to current reasoning LLMs and discuss three future research directions. In conclusion, we argue for properly adopting reasoning LLMs for financial AI research.

2 Preliminaries of DeepSeek-R1

In this section, we briefly introduce some preliminaries of reasoning LLMs, which are divided into three parts: the concept of CoTs, the DeepSeek-R1 training process, and DeepSeek-R1 distillation exploration.

2.1 Concept of CoTs

Following the emergence of GPT-3, researchers recognized the latent reasoning capabilities (Webb et al., 2023) inside LLMs, despite unreliability in some tasks. Specifically, it was discovered that even simple prompt engineering, such as adding "think step by step" (Kojima et al., 2022), could guide LLMs to generate human-like sequential reasoning texts and improve their performance for reasoning tasks. Fig. 1 shows an example where DeepSeek-V3 (Liu AX et al., 2024) is prompted to raise and answer a financial question with and without CoT. The observation that prompting LLMs can elicit step-by-step reasoning has sparked a growing research interest in how to systematically generate high-quality CoT. Subsequent work has explored various structured forms of CoT, including self-consistent CoT (Wei et al., 2022), tree-of-thought (Yao et al., 2023), and graph-of-thought (Besta et al., 2024). According to OpenAI-o1's technology report (Jaech et al., 2024), OpenAI researchers integrated CoT reasoning deeply into its training process, encouraging their model to make longer contemplation of CoT sequences before giving final answers through RL (Bai YT et al., 2022; Lightman et al., 2024; Zelikman et al., 2024). This advancement allows OpenAI-o1 to tackle more-complex reasoning tasks and provide clearer insights into its problem-solving processes, surpassing previous models such as GPT-4 in multistep reasoning tasks.

2.2 DeepSeek-R1 training process

To learn to generate accurate step-by-step reasoning CoTs, DeepSeek-R1 was trained under four carefully designed stages (Fig. 2a):

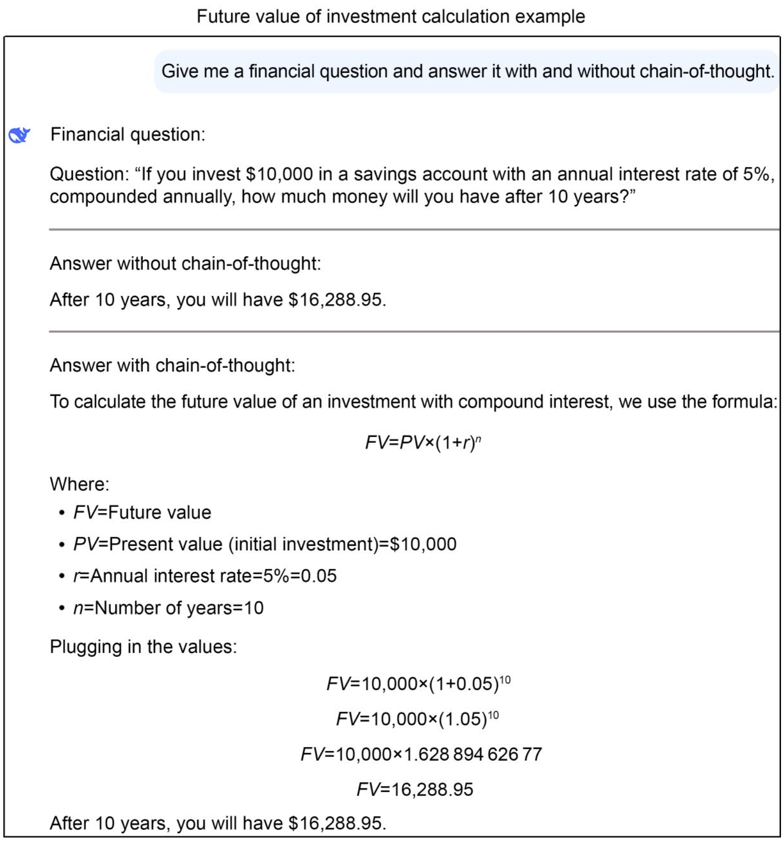

Fig. 1 Prompting DeepSeek-V3 to raise and answer a financial question with and without chain-of-thought

1. Cold-start supervised fine-tuning (SFT): Thousands of handcrafted high-quality CoTs were collected to train DeepSeek-V3 noted as M_0, a powerful LLM with 671B parameters, and model M_1 was obtained as a cold-start point.

2. Reasoning-focused RL: The DeepSeek team proposed a large-scale RL training strategy, whereby the group relative policy optimization (GRPO) algorithm (Shao et al., 2024) and rule-based reward models are integrated, to guide $M_1 \rightarrow M_2$ on reasoning-focused tasks. As a result, M_2 can learn how to perform multi-step thinking processes on coding, mathematics, science, and logic reasoning.

3. Rejection sampling and SFT: The generated high-quality CoTs ($n=6\times10^5$) were sampled and refined from the responses of M_2, with additional data on more reasoning tasks. These data together with 2×10^5 non-reasoning data, including writing, factual question answering, and role-playing tasks, were used to supervise the $M_2 \rightarrow M_3$ training process.

4. Final RL alignment: M_3 further went through a secondary RL stage to ensure its alignment to human preferences for helpfulness and safety, to reach DeepSeek-R1 M_4.

Given that the DeepSeek founder has a background as a quant investing researcher, it is reasonable to assume that the DeepSeek team incorporated finance-related reasoning data at stages 2 and 3. This addition is of two-fold significance. First, financial knowledge, being highly structured (Lusardi and Mitchell, 2014), can enhance the model's overall reasoning capabilities. The well-organized nature of financial concepts and rules provides a solid framework for the model to learn and improve its logical thinking. Second, as a result, the model is enabled to perform exceptionally well on financial reasoning-type questions. Therefore, DeepSeek-R1 performs well in our following financial QA experiments. It can also be concluded that high-quality CoTs play an essential role in the success of DeepSeek-R1. Note that the DeepSeek team has open-sourced the weight of DeepSeek-R1 M_4 (the model obtained after the fourth training stage), but not the training data.

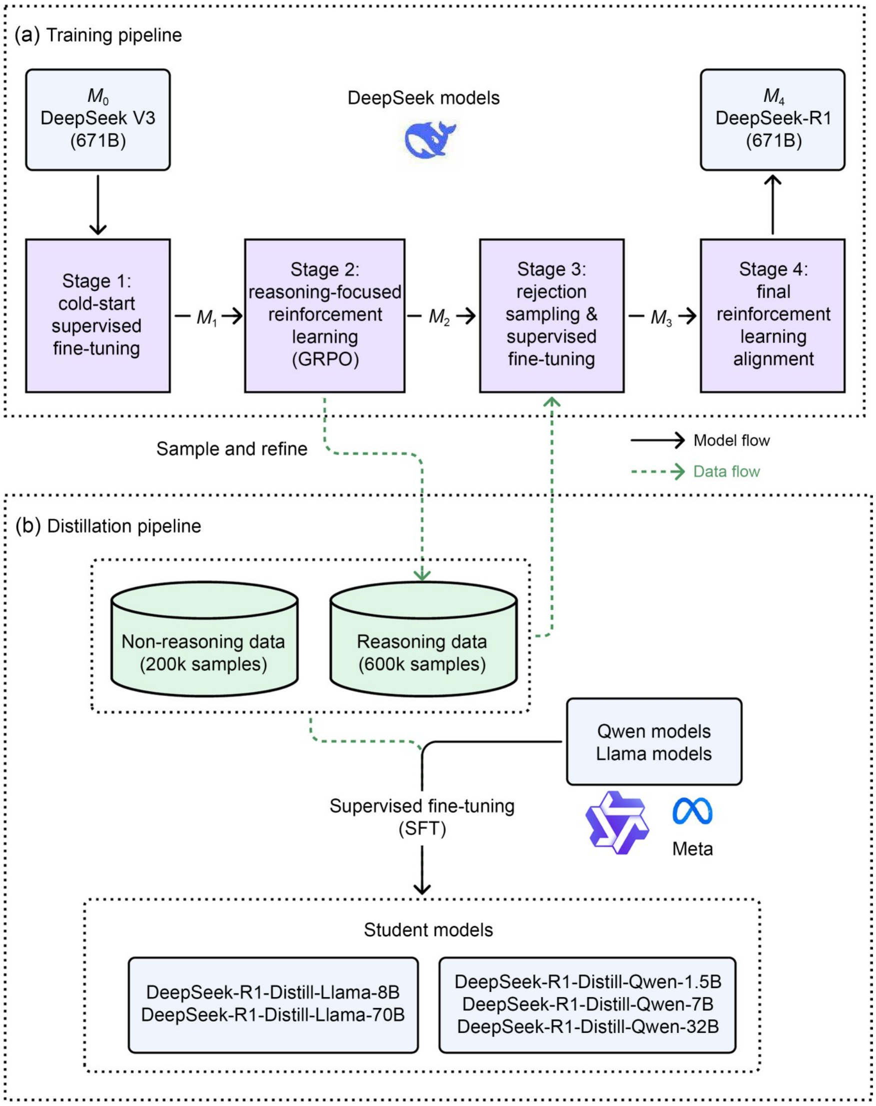

Fig. 2 Overview of the DeepSeek-R1 training pipeline: (a) four-stage training steps from DeepSeek V3 to DeepSeek-R1, incorporating group relative policy optimization (GRPO)-based reinforcement learning and alignment; (b) knowledge distillation pipeline using reasoning and non-reasoning data to train compact student models via supervised fine-tuning

2.3 DeepSeek-R1 distillation exploration

According to the training process, DeepSeek-R1 shares the same model size as DeepSeek-V3 with 671B parameters, which is not suitable for resource-constrained conditions. To equip smaller LLMs with reasoning ability, DeepSeek researchers explored to perform SFT on 1.5B–70B Qwen (Bai JZ et al., 2023) and Llama (Touvron et al., 2023) LLMs on the 800k samples in stage 3 (Fig. 2b), obtaining models including DeepSeek-R1-Distill-Qwen-32B and DeepSeek-R1-Distill-Llama-8B. The DeepSeek team has shown that distilling knowledge from these carefully sampled data is effective, and that smaller LLMs can achieve competitive results on math and coding tasks, compared to DeepSeek-R1.

Note that these student models are actually distilled from M_2, as described above, rather than from M_4. DeepSeek's exploration is an early trial to demonstrate the effectiveness of knowledge distillation (Hinton et al., 2015). We believe that exploring the ways to obtain better student models from DeepSeek-R1 M_4 will be a promising future research direction.

3 Benchmarking reasoning LLMs on financial question answering

In this section, we benchmark DeepSeek-R1 and its distilled student models on two public financial QA datasets with common LLM baselines, including DeepSeek-V3 (Liu AX et al., 2024), OpenAI-o1-preview (Jaech et al., 2024), and OpenAI GPT-4 (Achiam et al., 2023), to compare their performances in the finance area.

3.1 Datasets and evaluation protocols

1. Datasets: FinEval (Guo X et al., 2023) contains a total of 1321 single-choice QA pairs, which are categorized into four primary types: finance, economy, accounting, and certificate. Note that FinEval includes 4661 questions in total, where the answers to 3340 questions are not released. FinanceIQ (Zhang XY and Yang, 2023) includes a total of 7173 single-choice QA pairs from 10 subjects, such as the China Insurance Certification Exam, the US Certified Public Accountant Exam, and financial mathematics exams.

2. Evaluation protocols: All the LLMs are prompted to give final answers with JSON format CoTs and answers using the prompting template as shown in Fig. 3. It is noted that DeepSeek-R1 answers questions with a long reasoning content with the "<think>...<\think>" tag. We report the accuracy on two datasets, where accuracy is the ratio of the number of correctly answered questions to the total number of questions.

3.2 Main results

The experimental results are summarized in Table 1. We can find that reasoning LLMs DeepSeek-R1 and OpenAI-o1-preview significantly outperform their predecessor general LLMs DeepSeek-V3 and GPT-4 respectively, indicating that answering financial questions needs advanced reasoning ability. Furthermore, it can be observed that DeepSeek-R1 and DeepSeek-V3 perform better than OpenAI-o1-preview and GPT-4 respectively by a significant margin. We assume that this is because there exist data biases in the datasets used, whereby most financial questions are sampled from Chinese finance-related examinations and OpenAI-o1/GPT-4 lacks Chinese financial knowledge. We can also find that DeepSeek-R1-Distill-Qwen-32B has very competitive overall performance on these two datasets, considering that its learning parameters (32B) are much fewer than those in DeepSeek-V3/R1 (671B). With 7B learning parameters, which are much fewer than those in other models, DeepSeek-R1-Distill-Qwen-7B cannot achieve satisfactory results.

In addition to reporting the overall accuracy on FinEval and FinanceIQ, we specifically highlight the performance of the Financial Mathematics subset of FinanceIQ, as all models achieve their lowest scores in this category compared to other subjects. It can be noted that DeepSeek-R1 shows a more pronounced performance advantage in financial mathematics over other models. This observation further underscores the importance of specialized knowledge and advanced reasoning capabilities in handling complex financial calculations.

We present visualizations of two DeepSeek-R1 responses to financial mathematics questions in Fig. 3. DeepSeek-R1 can perform precise calculations and handle basic compound interest problems in finance. Nevertheless, because of its restricted advanced financial knowledge, it may be confused by irrelevant conditions in certain complex problems.

4 Discussion

4.1 Opportunities for financial applications of DeepSeek-R1

Based on the experimental results, DeepSeek-R1 is able to play the following three roles with its current financial knowledge:

1. Financial text processor: DeepSeek-R1 is capable of analyzing financial news, company announcements, and other textual contents in a more structured manner through logical reasoning. In our experiments, DeepSeek-R1 achieves an accuracy of more than 80% on questions related to the judgment and interpretation of Chinese economic laws and regulations. Therefore, DeepSeek-R1 can serve as a financial text processor to help financial officers in scenarios such as financial meeting summary, research report analysis, and compliance review.

2. Financial numerical analysis aide: DeepSeek-R1 significantly outperforms other models in FinanceIQ's Financial Mathematics category, as shown in Table 1. Accordingly, DeepSeek-R1 can provide users with

Question

甲公司年初累计未分配利润1000万元，本年净利润500万元。按法律规定，净利润要提取10%的盈余公积金。预计明年需要进行长期投资800万元，公司的目标资本结构是债务资本占40%、权益资本占60%。公司采用剩余股利政策，本年应分配的股利是（）万元。[Company A had an accumulated undistributed profit of 10 million yuan at the beginning of the year and a net profit of 5 million yuan this year. According to the law, 10% of the net profit should be set aside as surplus reserve. It is estimated that the company will need to make a long-term investment of 8 million yuan next year. The company's target capital structure is 40% debt capital and 60% equity capital. The company follows a residual dividend policy. The dividends to be distributed this year are ( ) *10 thousand yuan.]

A. 480 B. 0 C. 20 D.520

你需要一步一步思考，输出答案和思考过程，遵循如下JSON格式。[Please think step by step and provide the reasoning behind selecting the correct answer as the output. Your output should follow the JSON format.]

{
"think":你的思考过程，[Your step-by-step reasoning process]
"answer":你从A, B, C, D中选择的正确答案，[The correct choice from A, B, C, D]
}

Answer：C

DeepSeek-R1

<think>\nOK, I am now going to solve this accounting question. . ... (>1000 characters) ... In summary, the correct answer is Option B.<\think>

{"think":"首先，甲公司需要从上一财年的净利润500万元中提取10%的法定盈余公积金，即50万元，剩余450万元。[First, Company A needs to set aside 10% of the net profit from the previous fiscal year, which is 5 million yuan, as the statutory surplus reserve, that is, 0.5 million yuan, leaving 4.5 million yuan.]

根据剩余股利政策，公司需优先满足投资所需的权益资本。本年度计划增加800万元长期资本，权益部分为800×60%=480万元。[The company needs to prioritize meeting the equity capital required for investment. The company plans to increase long-term capital by 8 million yuan this year, with the equity portion being 8 million yuan × 60% = 4.8 million yuan]

由于当年的可分配利润（450万元）不足以满足权益需求（480万元），公司需将全部450万元用于投资，剩余股利为0。"[Since the distributable profit for the year (4.5 million yuan) is not sufficient to meet the equity requirement (4.8 million yuan), the company needs to use the entire 4.5 million yuan for investment, and the residual dividend is 0.]

"Answer"："B"}

Comment: **该回答考虑了题目中给出的所有条件，但它忽略了在剩余股利政策下，只需要考虑投资资本的成本，不需要考虑盈余公积。因此，它多扣减了一些不必要的项目。由于缺乏特定领域的金融知识，模型回答错误。实际上，我们使用净利润500万元减去权益资本成本480万元，得到的20万元即为股利，正确答案为C选项。**[The response takes into account all the conditions given in the question, but it overlooks the fact that **under the residual dividend policy, only the cost of investment capital needs to be considered**, and there is no need to consider the surplus reserve. Therefore, it deducts more items than necessary. The lack of specific financial knowledge leads to the **incorrect answer.** Actually, **we subtract the cost of equity capital, which is 4.8 million yuan, from the net profit of 5 million yuan. The resulting 200,000 yuan is the dividend**. So the correct answer is Option C.]

Question

A打算从一家企业那里借入10000元，借款的期限为二年。第一年的年实际贴现率为6%，而第二年则是以8%的年名义贴现率进行两次贴息。请问，A在年初时实际能借到的金额是多少元？ [A plans to borrow 10,000 yuan from a company for a period of two years. The actual annual discount rate for the first year is 6%, while in the second year, the discount is applied twice based on an annual nominal discount rate of 8%. What is the actual amount that A can borrow at the beginning of the year? ( ) yuan.]

A:9665 B:9660 C:8663 D:8667

你需要一步一步思考，输出答案和思考过程，遵循如下JSON格式。[Please think step by step and provide the reasoning behind selecting the correct answer as the output. Your output should follow the JSON format.]

{
"think":你的思考过程，[Your step-by-step reasoning process]
"answer":你从A, B, C, D中选择的正确答案，[The correct choice from A, B, C, D]
}

Answer：C

DeepSeek-R1

<think>\nOK, I am now going to solve this financial-math question. . ... (>1000 characters) ... In summary, the correct answer is Option C.<\think>

{"think":"借款两年后需偿还10000元。[The loan must be repaid in full after two years with a total repayment amount of 10,000 yuan.]

第二年贴现计算：名义贴现率8%分两次（每半年一次），每次贴现率4%。第二年末的10000元贴现到第二年初的现值为：[Discount Calculation for the Second Year: The nominal discount rate of 8% is applied twice (every six months), with each discount rate being 4%. The present value of 10,000 yuan at the end of the second year, discounted to the beginning of the second year, is]

$10000 \times (1-0.04)^2 = 9216$。

第一年贴现计算：将第二年初的9216元以第一年实际贴现率6%贴现到当前。[Discount Calculation for the First Year: Discount the 9,216 yuan at the beginning of the second year back to the present using the actual discount rate of 6% for the first year:]

$9216 \times (1-0.06) = 9216 \times 0.94 = 8663.04$"

"Answer"："C"}

Comment: **该回答将问题拆解为具体的计算步骤，并最终得出了正确答案。**[The response breaks down the problem into specific calculation steps and ultimately arrives at the **correct answer**.]

Fig. 3 Two cases of DeepSeek-R1's answers to finance math questions. The original Chinese and translated English contents are shown for readability. DeepSeek-R1's reasoning contents (>1000 characters) are in gray and shortened for space limitation. The same prompting template is used for all LLMs

Table 1 Benchmarking different LLMs on two financial QA datasets with accuracy in terms of correctly answering questions

Model	Accuracy (%)		
	FinEval	FinanceIQ	FinanceIQ's Financial Math
DeepSeek-V3	82.20	73.46	54.84
DeepSeek-R1	**91.52**	**86.35**	**82.80**
DeepSeek-R1-Distill-Qwen-32B	82.73	77.19	56.99
DeepSeek-R1-Distill-Qwen-7B	51.17	45.63	18.28
GPT-4	70.03	63.40	37.50
OpenAI-o1-preview	78.11	79.20	72.73

Best results are in bold

numerical analysis insights that have a certain level of credibility in various financial services, serving as an analysis assistant. However, note that while the calculations are deemed credible, they are not infallible and should be used with due caution.

3. Financial educator: Considering that the text questions were collected from multiple qualification examinations for various financial professions, and because DeepSeek-R1 can provide not only highly accurate choices but also detailed problem-solving analysis steps, DeepSeek-R1 can act as a talented educator, tutoring beginners in mastering relevant financial knowledge. However, during this process, the involvement of seasoned professionals is still required to identify any errors in DeepSeek-R1's responses.

4.2 Limitations of DeepSeek-R1 in financial applications

While DeepSeek-R1 has achieved satisfactory performance, it is essential to recognize that there is still a gap between the examined questions and real-world financial practices. We identify the following four limitations in applying DeepSeek-R1 to the finance domain:

1. Lack of evaluation on large-scale financial reasoning datasets: Our experiments are based on small-scale, relatively simple single-choice question datasets as an early step, and DeepSeek-R1 has not been comprehensively evaluated for financial fields. It is necessary to establish more large-scale datasets with a greater diversity of financial question types and more advanced reasoning application scenarios. Such datasets can be used to not only evaluate the performance of reasoning LLMs but also lay the data foundation for further improving the performance of reasoning LLMs in financial domains.

2. Reduced but still-existing hallucination: The financial sector is an area under high regulation, and even a minor error can have far-reaching consequences. Although DeepSeek-R1 can reduce the impact of hallucinations by demonstrating step-by-step reasoning process, hallucination risks still persist and are more difficult to detect. There is a need to explore more-advanced strategies to enhance the interpretability of DeepSeek-R1's black-box nature and reduce hallucinations, such as combining neural and symbolic operations (Chervonyi et al., 2025) and developing trustworthy LLMs (Liu Y et al., 2023).

3. Lack of model collaboration: For elementary financial text analysis tasks, such as corporate entity extraction, strong reasoning capabilities are not always necessary. Generation of excessively long thought chains by DeepSeek-R1 may instead reduce processing efficiency, while DeepSeek-V3, which directly provides answers, may be more cost-effective. Thus, DeepSeek-R1 needs to learn more from humans, who can seamlessly switch between fast (system 1) and slow (system 2) thinking (Hua and Zhang, 2022) and thus strike a balance through model collaboration.

4. Lack of multimodality ability: Beyond text-based data, the financial services domain requires handling multimodal data, including charts, images, audio, and price time-series data. At present, DeepSeek-R1 is limited to the text modality. Expanding its capabilities from text-centric domains to multimodal scenarios remains for exploration.

4.3 Future directions for DeepSeek-R1 in financial applications

DeepSeek-R1 has already led the discussion among AI practitioners across various industries. From the perspective of FinTech, here we put forward three potential research directions related to DeepSeek-R1:

1. Federated financial reasoning LLMs: Considering the DeepSeek-R1 training process, the scaling law (Kaplan et al., 2020) of high-quality reasoning data still holds. However, building large-scale financial datasets is challenging, considering privacy laws and financial regulations. To address this issue, the paradigm of federated learning (Fan et al., 2023; Kuang et al., 2024) can be adopted for training financial reasoning LLMs, where data privacy can be preserved. An implementation of federated financial LLM training frameworks is illustrated in Fig. 4. Moreover, protecting the high-quality reasoning CoT data generated by financial LLMs is an important research topic for federated financial reasoning LLMs. Exploring how to implicitly inject "copyright-protection watermark" information into the reasoning CoT is a worthy research direction.

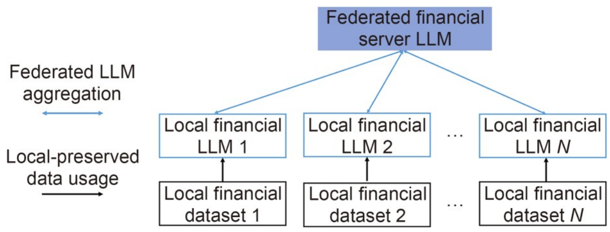

Fig. 4 An implementation of federated financial LLM training frameworks

2. Multimodel collaborative financial AI: Currently, DeepSeek-R1 requires extensive CoTs to solve every given problem. It is necessary to consider integrating fast thinking to more efficiently address problems in financial scenarios. Therefore, future financial AI should be constructed through the collaboration of fast- and slow-thinking models, and there have been some early attempts (Miao et al., 2024; Sun et al., 2024). In addition to this paradigm combination, considering that DeepSeek-R1 has explored distilling easy-to-deploy student models with certain reasoning capabilities, methods for the cooperation between large and small language models to achieve a more computation-friendly and low-carbon financial AI need to be explored.

3. Multimodality financial agent: Building of advanced agent systems for financial services requires handling data from multiple modalities. Key research directions include developing multimodality reasoning large models that can understand and generate different modalities of financial data. Previous studies (Daiya and Lin, 2021; Zhang WT et al., 2024) have demonstrated that integrating multimodality features can further enhance performance compared to relying solely on single-modality text data.

5 Conclusions

In conclusion, DeepSeek-R1 is set to inspire in-depth research and diverse applications in financial services. This paper details DeepSeek-R1's technology, benchmarks its performance on two financial QA datasets, and discusses its opportunities and limitations. As an early effort to spark interest in integrating reasoning LLMs into financial AI, however, our study has the following limitation: it has been tested only on single-choice questions with insufficient model comparisons. We leave the upgradation of this research for our future work and also as open problems to the research community. This work proves DeepSeek-R1's potential in finance, and we advocate proper use of reasoning LLMs and suggest future financial AI research directions.

Contributors

Shuoling LIU, Liyuan CHEN, and Jiangpeng YAN designed the research. Shuoling LIU, Jiangpeng YAN, Yuhang JIANG, and Xiaoyu WANG processed the data. Shuoling LIU, Jiangpeng YAN, Yuhang JIANG, and Xiaoyu WANG drafted the paper. Liyuan CHEN and Xiu LI helped organize the paper. Xiu LI and Qiang YANG revised and finalized the paper.

Conflict of interest

Shuoling LIU, Liyuan CHEN, Xiu LI, and Qiang YANG are guest editors of the Special Feature on Theories and Applications of Financial Large Models of *Frontiers of Information Technology & Electronic Engineering*; they were not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Achiam J, Adler S, Agarwal S, et al., 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774

Bai JZ, Bai S, Chu YF, et al., 2023. Qwen technical report. https://arxiv.org/abs/2309.16609

Bai YT, Jones A, Ndousse K, et al., 2022. Training a helpful and harmless assistant with reinforcement learning from

human feedback. https://arxiv.org/abs/2204.05862
Besta M, Blach N, Kubicek A, et al., 2024. Graph of thoughts: solving elaborate problems with large language models. Proc 38th AAAI Conf on Artificial Intelligence, p.17682-17690. https://doi.org/10.1609/aaai.v38i16.29720
Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Proc 34th Int Conf on Neural Information Processing Systems, Article 159.
Chang YP, Wang X, Wang JD, et al., 2024. A survey on evaluation of large language models. *ACM Trans Intell Syst Technol*, 15(3):39. https://doi.org/10.1145/3641289
Chervonyi Y, Trinh TH, Olšák M, et al., 2025. Gold-medalist performance in solving Olympiad geometry with AlphaGeometry2. https://arxiv.org/abs/2502.03544
Daiya D, Lin C, 2021. Stock movement prediction and portfolio management via multimodal learning with Transformer. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.3305-3309. https://doi.org/10.1109/ICASSP39728.2021.9414893
Fan T, Kang Y, Ma GQ, et al., 2023. FATE-LLM: an industrial grade federated learning framework for large language models. https://arxiv.org/abs/2310.10049
Gemini Team of Google, 2023. Gemini: a family of highly capable multimodal models. https://arxiv.org/abs/2312.11805
Guo DY, Yang DJ, Zhang HW, et al., 2025. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. https://arxiv.org/abs/2501.12948
Guo X, Xia HT, Liu ZW, et al., 2023. FinEval: a Chinese financial domain knowledge evaluation benchmark for large language models. https://arxiv.org/abs/2308.09975
Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531
Hua WY, Zhang YF, 2022. System 1+system 2=better world: neural-symbolic chain of logic reasoning. Proc Findings of the Association for Computational Linguistics, p.601-612. https://doi.org/10.18653/v1/2022.findings-emnlp.42
Huang L, Yu WJ, Ma WT, et al., 2025. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans Inform Syst*, 43(2):42. https://doi.org/10.1145/3703155
Jaech A, Kalai A, Lerer A, et al., 2024. OpenAI o1 system card. https://arxiv.org/abs/2412.16720
Kaplan J, McCandlish S, Henighan T, et al., 2020. Scaling laws for neural language models. https://arxiv.org/abs/2001.08361
Kojima T, Gu SS, Reid M, et al., 2022. Large language models are zero-shot reasoners. Proc 36th Int Conf on Neural Information Processing Systems, Article 1613.
Kuang WR, Qian BC, Li ZT, et al., 2024. FederatedScope-LLM: a comprehensive package for fine-tuning large language models in federated learning. Proc 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.5260-5271. https://doi.org/10.1145/3637528.3671573
Lightman H, Kosaraju V, Burda Y, et al., 2024. Let's verify step by step. Proc 12th Int Conf on Learning Representations, p.1-24.
Liu AX, Feng B, Xue B, et al., 2024. DeepSeek-V3 technical report. https://arxiv.org/abs/2412.19437
Liu Y, Yao YS, Ton JF, et al., 2023. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. https://arxiv.org/abs/2308.05374
Lusardi A, Mitchell OS, 2014. The economic importance of financial literacy: theory and evidence. *J Econ Liter*, 52(1): 5-44. https://doi.org/10.1257/jel.52.1.5
Miao XP, Oliaro G, Zhang ZH, et al., 2024. SpecInfer: accelerating large language model serving with tree-based speculative inference and verification. Proc 29th ACM Int Conf on Architectural Support for Programming Languages and Operating Systems, p.932-949. https://doi.org/10.1145/3620666.3651335
Peng YF, Malin BA, Rousseau JF, et al., 2025. From GPT to DeepSeek: significant gaps remain in realizing AI in healthcare. *J Biomed Inform*, 163:104791. https://doi.org/10.1016/j.jbi.2025.104791
Qian LF, Zhou WP, Wang Y, et al., 2025. Fino1: on the transferability of reasoning enhanced LLMs to finance. https://arxiv.org/abs/2502.08127
Shao ZH, Wang PY, Zhu QH, et al., 2024. DeepSeekMath: pushing the limits of mathematical reasoning in open language models. https://arxiv.org/abs/2402.03300
Sun GY, Jin MY, Wang ZT, et al., 2024. Visual agents as fast and slow thinkers. https://arxiv.org/abs/2408.08862
Touvron H, Martin L, Stone K, et al., 2023. Llama 2: open foundation and fine-tuned chat models. https://arxiv.org/abs/2307.09288
Vaswani A, Shazeer N, Parmar N, et al., 2017. Attention is all you need. Proc 31st Int Conf on Neural Information Processing Systems, p.6000-6010.
Webb T, Holyoak KJ, Lu HJ, 2023. Emergent analogical reasoning in large language models. *Nat Human Behav*, 7(9): 1526-1541. https://doi.org/10.1038/s41562-023-01659-w
Wei J, Wang XZ, Schuurmans D, et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Proc 36th Int Conf on Neural Information Processing Systems, Article 1800.
Wu SJ, Irsoy O, Lu S, et al., 2023. BloombergGPT: a large language model for finance. https://arxiv.org/abs/2303.17564
Yao SY, Yu D, Zhao J, et al., 2023. Tree of thoughts: deliberate problem solving with large language models. Proc 37th Int Conf on Neural Information Processing Systems, Article 517.
Zelikman E, Harik G, Shao Y, et al., 2024. Quiet-STaR: language models can teach themselves to think before speaking. https://arxiv.org/abs/2403.09629
Zhang WT, Zhao LX, Xia HC, et al., 2024. A multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist. Proc 30th ACM SIGKDD Conf on Knowledge Discovery and Data Mining, p.4314-4325. https://doi.org/10.1145/3637528.3671801
Zhang XY, Yang Q, 2023. XuanYuan 2.0: a large Chinese financial chat model with hundreds of billions parameters. Proc 32nd ACM Int Conf on Information and Knowledge Management, p.4435-4439. https://doi.org/10.1145/3583780.3615285
Zhou J, Ke P, Qiu XP, et al., 2024. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, 25(1):6-11. https://doi.org/10.1631/FITEE.2300089