CLER: Improving Multimodal Financial Reasoning by Cross-MLLM Error Reflection

Shuangyan Deng[1], **Zhongsheng Wang**[1], **Rui Mao**[2], **Ciprian Doru Giurcăneanu**[1], **Jiamou Liu**[1]
[1]University of Auckland, Auckland, New Zealand
[2]Nanyang Technological University, Singapore
sden118@aucklanduni.ac.nz, {zhongsheng.wang, c.giurcaneanu, jiamou.liu}@auckland.ac.nz, rui.mao@ntu.edu.sg

Abstract

Recent advances in Multimodal Large Language Models (MLLMs) have enabled joint reasoning over financial textual and visual inputs. However, they still struggle with financial terminology, logical consistency, and numerical computations. Moreover, while commercial large models perform well on reasoning tasks, their high inference costs limit their scalable usage in real world financial applications. We thus propose a cost-effective framework, CLER, that combines contrastive retrieval with step-wise reflection to improve reasoning performance. Also, the reasoning cost is only generated in the test stage when using commercial large models. CLER leverages *FinErrorSet*, a dataset of 8,000+ mistake-correction pairs from diverse open-source MLLMs. A fine-grained retriever is trained to identify structurally relevant errors for self-correction through individual reflection. Experiments on three benchmarks show that CLER consistently outperforms other baselines. To our knowledge, CLER is the first framework to use cross-model errors for financial reasoning.

Code — https://github.com/Jennie-Deng/CLER

Introduction

A smart man learns from his own mistakes,
a wise man learns from the mistakes of others.
— Otto von Bismarck

Financial reasoning improves the accuracy and interpretability of financial analysis by providing a structured and transparent thinking process. Large Language Models (LLMs) have demonstrated strong performance on financial reasoning tasks (Zhu et al. 2021; Zhao et al. 2024b; Du et al. 2025a). However, the knowledge-intensive and multimodal nature of financial data makes the domain particularly suitable for Multimodal LLMs (MLLMs). Applications like report analysis (Zhao et al. 2024a), risk forecasting (Sawhney et al. 2020), and market analysis (Liu et al. 2024) particularly benefit from models capable of jointly processing text and visual data. However, recent benchmarks (Luo et al. 2025; Rangapur et al. 2025a; Gan et al. 2024; Xue et al. 2024; Deng et al. 2025) reveal that mainstream MLLMs (e.g., GPT-4o, Claude-3.5) still perform poorly on financial reasoning, with accuracy often below 60%. These evaluations highlight persistent challenges in contextual integration, logical reasoning, and numerical computation. Moreover, closed-source models, despite generally exhibiting stronger performance, incur high inference costs due to the token explosion that occurs during image processing. Specifically, as the image resolution increases, the number of vision tokens rises quadratically, significantly escalating computational complexity and inference latency (Zhao et al. 2025; Li et al. 2024b; Guo et al. 2024).

Learning from mistakes, particularly those made by others, is a cognitively grounded and empirically supported strategy for improving reasoning (Mao et al. 2024). *Social Cognitive Theory* (Bandura 1977) suggests that individuals effectively acquire knowledge and behavioral strategies by observing others, a process known as *vicarious learning*. Learning from others' mistakes is also a smarter strategy for mitigating the risk of significant losses. Studies on erroneous example-based learning (Van Gog and Rummel 2010; Durkin and Rittle-Johnson 2012) further support this view, showing that exposure to incorrect examples promotes deeper understanding by prompting learners to reflect, diagnose, and revise errors. This aligns with the notion of *negative knowledge* (Tulis 2013), which emphasizes learning what not to do in complex tasks. Recent studies show that negative examples can guide LLMs to reflect and avoid previous errors (An et al. 2024; Tong et al. 2024; Zhou et al. 2024; Zhang et al. 2024; Ranga et al. 2024).

Inspired by these cognitive foundations, we investigate whether MLLMs can enhance financial reasoning by learning from one another's mistakes. Specifically, we examine how open-source MLLMs can provide informative negative examples to help closed-source models enhance both accuracy and efficiency in financial multimodal tasks. This raises our central question: *To what extent can MLLMs benefit from cross-model error reflection compared to naive reasoning?*

However, applying the principle of learning from mistakes across different MLLMs presents several non-trivial challenges. First, identifying and extracting meaningful error instances from model-generated reasoning processes is challenging due to the knowledge-intensive and multimodal nature of financial reasoning tasks, which involve advanced financial concepts, intricate logical and numerical reasoning, and high context dependency. Second, transferring negative

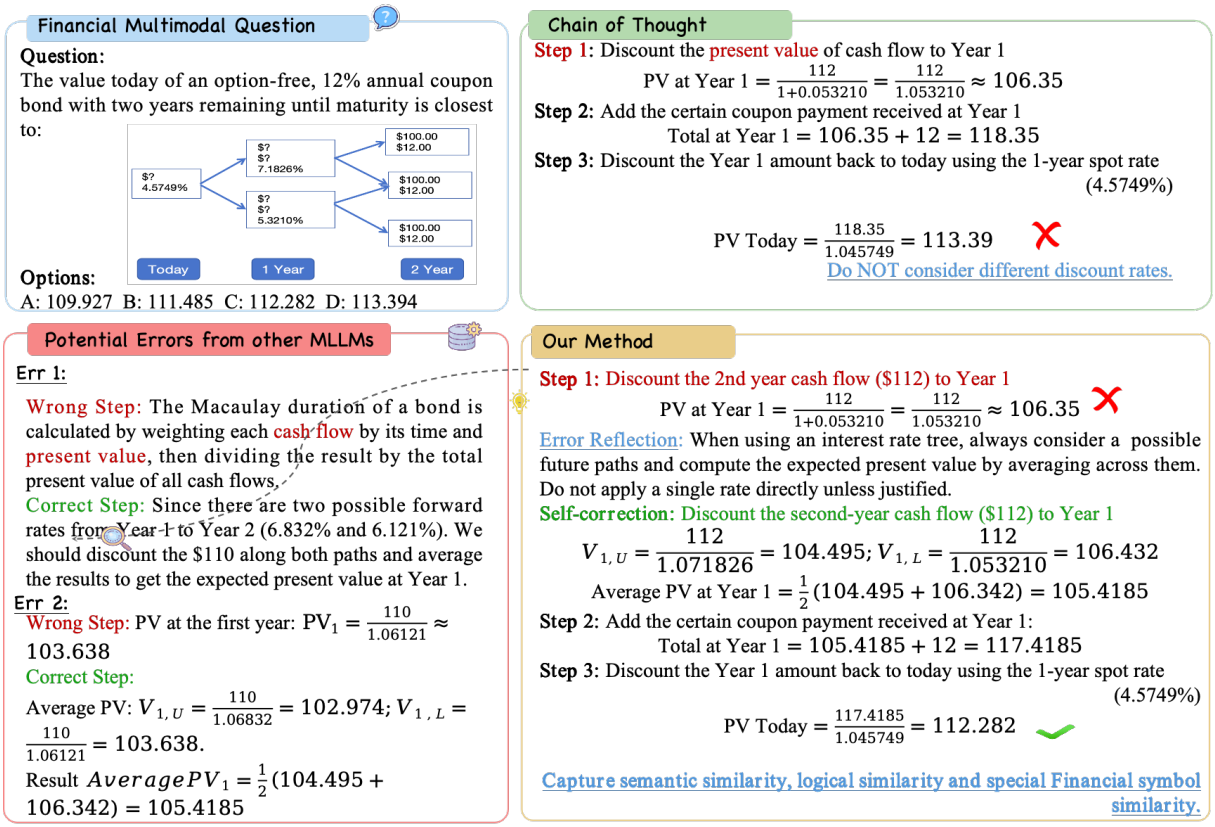

Figure 1: An illustrative example of financial multimodal reasoning via CoT and CLER.

knowledge between heterogeneous MLLMs is non-trivial because their reasoning styles often differ substantially, making direct reuse of error patterns ineffective. Third, in multi-step reasoning tasks, general errors (e.g., applying incorrect formulas or misinterpreting financial logic) are often overly broad and nonspecific, offering little actionable guidance for targeted model improvement.

To address these challenges, we propose Contrastive Learning for Error Retrieval (CLER) – the first framework that systematically leverages cross-model errors from cost-effective open-source MLLMs to conduct reflection, ultimately enhancing financial multimodal reasoning (Figure 1). CLER addresses these challenges by: (1) training a contrastive retrieval model that captures both semantic and structural similarities between reasoning steps to identify meaningful, task-specific error instances; (2) enabling individualized reflection by allowing each MLLM to adapt retrieved errors to its own reasoning style rather than directly transferring generalized error rules; and (3) retrieving fine-grained, step-level mistake-correction pairs, enabling targeted recognition of context-dependent errors (e.g., misapplying a formula at step 3 or omitting a negative sign at step 4). This fine-grained, context-aware error reflection facilitates more precise self-correction and robust improvements in multimodal financial reasoning.

The contributions are summarized as follows: 1) We propose **CLER**, a novel framework, which systematically leverages reasoning errors made by cost-effective open-source MLLMs to enhance the reasoning performance of closed-source MLLMs. Extensive experiments demonstrate that CLER consistently outperforms baselines, achieving higher accuracy improvement. 2) We introduce and publicly release *FinErrorSet*, the first large-scale financial reasoning error dataset, consisting of over 8,000 mistake-correction pairs at the reasoning-step level. 3) We train a contrastive retrieval model on *FinErrorSet* to enable accurate retrieval of similar error patterns from different MLLMs, which can effectively retrieve semantically and structurally similar wrong steps from an external mistake database.

Related Work

Financial Multimodal Reasoning

In the financial sector, characterized by intricate decision-making processes, domain-specific expertise, and sophisticated analysis, LLMs have demonstrated substantial potential in enhancing financial reasoning and analytical tasks, such as investment decision-making (Chen et al. 2021; Wu et al. 2023; Zhang and Yang 2023; Yang, Liu, and Wang 2023; Xie et al. 2023; Zhao et al. 2023, 2024c; Du et al. 2024). Notably, recent evaluations indicate that advanced LLMs now surpass human performance on established financial reasoning datasets, including TAT-QA (Zhu et al. 2021), TabMWP (Zhu et al. 2021), FinQA (Chen et al.

2021), and DocMath (Zhao et al. 2023). Beyond text-based analysis, financial multimodal reasoning introduces additional complexity and potential for innovation. Generally, financial multimodal reasoning tasks can be categorized into two types: *knowledge-driven reasoning* and *perception-driven reasoning*. Knowledge-driven reasoning (Deng et al. 2025; Gan et al. 2024; Xue et al. 2024; Du et al. 2025b) requires MLLMs to leverage deep financial expertise, handle advanced financial terminologies, interpret nuanced contextual information, and analyze complex financial visual data such as charts, graphs, and tables. In contrast, perception-driven reasoning (Luo et al. 2025; Rangapur et al. 2025b; Gan et al. 2024; Bhatia et al. 2024) mainly assesses basic visual capabilities, requiring interpretation of simpler visual elements without extensive integration of textual context.

Despite the expectation that MLLMs approach human-level reasoning performance on multimodal tasks, empirical evaluations on financial multimodal benchmarks, such as FinMR (Deng et al. 2025), FAMMA (Xue et al. 2025), and FinMME (Luo et al. 2025), consistently demonstrate performance below 60%, indicating significant limitations. These findings highlight persistent challenges faced by MLLMs in effectively integrating complex textual and visual financial data, underscoring the critical need for further advancement in multimodal financial reasoning methodologies.

Learning from Mistakes

Researchers have recently proposed various methods for integrating negative or erroneous examples to enhance the performance of LLMs (An et al. 2024; Tong et al. 2024; Li et al. 2024a; Wang et al. 2024; Zhang et al. 2024; Sun et al. 2024; Jin et al. 2025). One prominent approach involves fine-tuning LLMs with negative data to guide them away from common errors (Mao et al. 2024). However, this strategy is impractical for closed-source models (e.g., GPT series and major MLLMs). An alternative to fine-tuning is to store historical mistakes in external databases and retrieve these examples dynamically to inform model reasoning. Methods such as Contrastive Chain-of-Thought (Chia et al. 2023), which explicitly provide both positive and negative reasoning examples within prompts, have been proposed. Additionally, recent techniques like LEAP (Zhang et al. 2024) and RICP (Sun et al. 2024) utilize generalized insights and principles derived from past errors to guide LLM reasoning.

However, directly applying these methods to multimodal financial reasoning tasks incurs prohibitively high token-processing costs due to the complexity and size of multimodal data. Consequently, a more cost-effective strategy involves learning from errors generated by open-source models, avoiding the resource-intensive demands of closed-source multimodal models. Moreover, rather than generalizing principles broadly, adopting a step-level individualized error learning mechanism could yield more precise and effective corrections tailored to each reasoning scenario.

CLER for Multimodal Financial Reasoning

Task Definition

We define multimodal financial reasoning as a structured reasoning task that requires the joint integration of textual and visual financial information. Formally, each instance is represented as: $X = (c, q, I, O)$, where c denotes the textual financial context, q is the reasoning question, $I = \{i_1, \ldots, i_s\}$ represents the associated visual inputs (e.g., stock charts or financial tables), and $O = \{o_1, \ldots, o_n\}$ is the set of candidate answers, with exactly one correct answer $A^* \in O$. Given an input instance X, the model is required to generate a step-wise reasoning trajectory $\hat{\mathcal{R}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_T\}$ that leads causally and interpretably to a predicted answer $\hat{A}$, with each step $\hat{r}_t$ grounded in the provided evidence. The optimization objective is to accurately predict the ground-truth answer A^* by minimizing the prediction error between $\hat{A}$ and A^*, while producing coherent reasoning trajectories that align with financial evidence, thereby ensuring both accuracy and interpretability.

CLER Framework Overview

Motivated by the principle of *learning from others' mistakes*, CLER aims to enhance the performance of financial multimodal reasoning tasks by enabling models to reflect on errors made by other models. As illustrated in Figure 2, CLER consists of three key stages: 1) constructing an error database (*FinErrorSet*) that records erroneous reasoning steps along with their corresponding corrections, 2) training a contrastive retriever that identifies semantically and structurally similar errors, and 3) facilitating model-specific reflection to promote effective self-correction. We will describe each module in detail below.

Error Database Construction

To facilitate fine-grained learning from diverse errors, we construct an error database $\mathcal{D}$, referred to as *FinErrorSet* (see Figure 2(a)). To be specific, we employ multiple open-source MLLMs, including DeepSeek-VL-2 (Wu et al. 2024), Qwen-VL-2.5 (QwenTeam 2025), and InternVL-3 (Chen et al. 2024), to generate reasoning chains $\mathcal{R}$ based on the FinMR (Deng et al. 2025) and FAMMA (Xue et al. 2024) training datasets. The corresponding reasoning chains are generated using the following prompt:

> **[System Input]:** You are a financial expert and are supposed to answer the given questions with options, context information, and images. Put the correct option (A, B, C, or D) in **[]**. e.g. Therefore, the correct option is **[B]**.
> **[User Input]:** Context: {context}; Question: {question}; Images: {images}; Options: {options}
> Let's think step by step to answer the given question.

Each generated reasoning chain $\mathcal{R}$ consists of multiple intermediate steps $(r_{i,j})$, where i denotes the step index and j represents the question index. Each step is then independently evaluated by the same models acting as step checkers, which compare each generated step against ground-truth annotations, in Figure 2(a). Once an erroneous step $r'_{i,j}$ is iden-

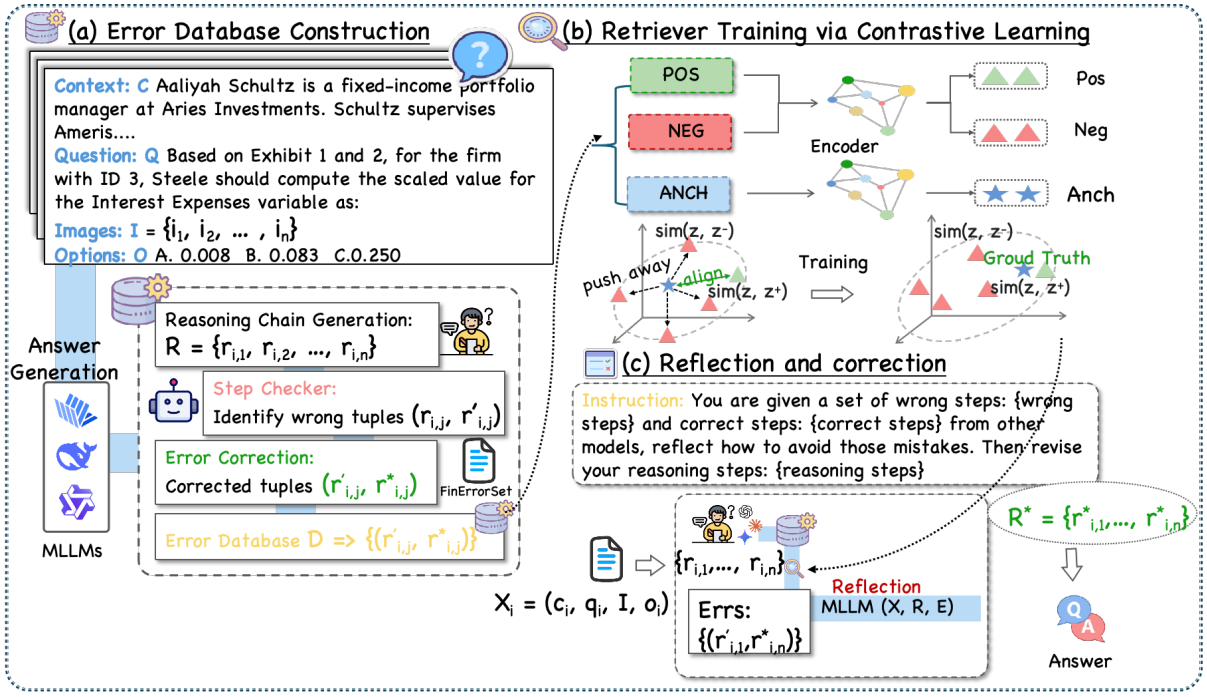

Figure 2: The framework of CLER includes: (a) Error Database Construction: all three open-source models are used to solve multimodal financial problems step by step, (b) A pretrained retriever via contrastive learning method, (c) Reflection and Correction: proprietary models firstly generate reasoning chains, then use pretrained retriever to find top-1 similar wrong step and correct step for each step form other models, finally refect whether they have same problems and conduct self-correction.

tified, it is revised to obtain the corrected version $r^*_{i,j}$, forming an error-correction pair $(r'_{i,j}, r^*_{i,j})$.

[System Input]: You are a step checker, you compare your wrong reasoning steps and correct explanations. Accordingly, revise your wrong reasoning steps.
[User Input]: Context: {context}; Question: {question}; Images: {images}; Options: {options}; a set of wrong steps: {wrong steps} and correct explanations: {explanations}.

	Number	Ratio
Total Questions	3258	–
Total Wrong Steps	8852	–
Reasoning Type		
* Expertise Reasoning	5138	58.0%
* Math Reasoning	3714	42.0%
Source Models		
* DeepSeek-VL-2	2668	30.1%
* InternVL-3	2689	30.4%
* Qwen-VL-2.5	3495	39.5%

Table 1. Statistics of the Error Database: *FinErrorSet*.

Formally, the final database can be represented as: $\mathcal{D} = \{(r'_{1,1}, r^*_{1,1}), ..., (r'_{i,j}, r^*_{i,j})\}$. Table 1 shows the statistics of *FinErrorSet*. It captures a diverse set of error types produced by different MLLMs, including numerical miscalculations, financial logic misunderstandings, and incorrect multimodal integration. This structured error database serves as the basis for our retrieval model, supporting fine-grained reflection and correction for other models.

Contrastive Learning for Error Retrieval

Different models exhibit distinct reasoning styles due to variations in their architectures, training datasets, internal mechanisms, and development methodologies. To robustly identify and retrieve meaningful reasoning errors from the meticulously constructed error database $\mathcal{D}$ (*FinErrorSet*), we employ a contrastive learning approach. Specifically, we train a contrastive retrieval model designed to accurately capture and retrieve reasoning errors that are semantically and structurally similar in nature. Figure 2(b) outlines our training approach, with detailed parameter settings provided in Algorithm 1. As previously noted, Figure 1 illustrates an example where the retriever successfully identifies a relevant erroneous reasoning step along with its corresponding correction, demonstrating the retriever's effectiveness in capturing fine-grained reasoning similarities.

Step-level Error Reflection and Correction

When presented with the same error cases, individuals often exhibit diverse interpretations and reasoning paths. Similarly, different LLMs demonstrate varying reflective behaviors when addressing the same erroneous reasoning. To avoid ineffective guidance resulting from the homogenization of principles, we propose a model-specific reflec-

Algorithm 1: Retriever Training via Contrastive Learning

Input: Error database $\mathcal{D} = \{(r'_{i,j}, r^*_{i,j})\}$, batch size $B = 128$, epochs $E = 5$, temperature $\tau = 0.07$, encoder $En(\cdot)$: Sup-SimCSE

Output: Trained encoder $En(\cdot)$

Initialize encoder $En(\cdot)$
for epoch $e = 1, 2, \ldots, E$ **do**
  Shuffle error database $\mathcal{D}$
  for each batch $\{(r'_i, r^*_i)\}_{i=1}^{B} \subseteq \mathcal{D}$ **do**
    Encode anchor and positive $\boldsymbol{e}_i = En(r'_i)$, $\boldsymbol{e}_i^+ = En(r^*_i)$
    $\mathcal{L}_{\text{batch}} \leftarrow 0$
    for $i = 1, \ldots, B$ **do**
      $\text{sim}_i^+ = \frac{\boldsymbol{e}_i \cdot \boldsymbol{e}_i^+}{\|\boldsymbol{e}_i\| \|\boldsymbol{e}_i^+\|}$
      $\mathcal{N}_i \leftarrow \left\{En(r'_{a,b}) \mid (a,b) \neq (i,j)\right\}$
      Compute $\text{sim}_i^- = \left\{\cos(\boldsymbol{e}_i, \boldsymbol{e}^-) \mid \boldsymbol{e}^- \in \mathcal{N}_i\right\}$
      $\text{denom}_i = \exp(\text{sim}_i^+/\tau) + \sum_{\text{sim} \in \text{sim}_i^-} \exp(\text{sim}/\tau)$
      $\mathcal{L}_i = -\log \frac{\exp(\text{sim}_i^+/\tau)}{\text{denom}_i}$
      $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}} + \mathcal{L}_i$
    end for
    $\mathcal{L}_{\text{batch}} \leftarrow \mathcal{L}_{\text{batch}}/B$
    Update $En(\cdot)$ via gradient descent on $\mathcal{L}_{\text{batch}}$
  end for
end for
return trained encoder $En(\cdot)$

tion strategy. In particular, we encourage models to engage in individual reflection, guided by referenced error cases, enabling them to autonomously refine their reasoning chains and improve their answers. In the inference stage, the MLLM generates a multi-step reasoning chain $\mathcal{R} = \{r_1, r_2, \ldots, r_T\}$. For each reasoning step $r_t \in \mathcal{R}$, the retrieval model queries the constructed error database $\mathcal{D}$ to identify the top-1 semantically and structurally similar erroneous step $r'_{i,j}$ and its corresponding correction $r^*_{i,j}$.

The MLLM then performs error reflection by analyzing and summarizing these retrieved pairs and generating a corrective insight $\mathcal{I}(r_t, r'_{i,j}, r^*_{i,j})$. This insight is applied to revise the current reasoning step via a reflection function: $\hat{r}_t = \text{Self-correction}(r_t, \mathcal{I}(r_t, r'_{i,j}, r^*_{i,j}))$, yielding the refined reasoning chain $\hat{\mathcal{R}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_T\}$. This reflection mechanism enables step-wise self-correction, allowing the MLLM to iteratively revise flawed reasoning steps and improve accuracy. Below is the reflection prompt template.

[System Input]: You are given a set of wrong steps: wrong steps and correct steps: correct steps from other models, and reflect on how to avoid those mistakes. Then revise your reasoning steps: reasoning steps
[User Input]: Context: {context}; Question: {question}; Images: {images}; Options: {options}.

Experiments

Experimental Setup

Benchmarks We evaluate our approach on three financial multimodal reasoning benchmarks, each requiring integration of domain-specific knowledge, textual comprehension, and visual perception, such as tables and charts: 1) **FinMR** (Deng et al. 2025): An expert-level benchmark with 3,200 questions that require complex financial reasoning. Each instance includes detailed textual context, one or more financial images, and expert-annotated explanations. 2) **FAMMA** (Xue et al. 2024): A multilingual financial multimodal reasoning benchmark designed to reflect college-level expertise. Each financial question is paired with one or more relevant images, requiring advanced domain knowledge to support sophisticated reasoning. 3) **MMMU-Finance** (Yue et al. 2024): A specialized subset of the open-domain Multidiscipline Multimodal Understanding benchmark, comprising 1,110 samples selected from Finance, Economics, and Accounting domains[1]. Notably, 95% of the visual inputs in this subset are financial tables.

Baselines We consider the following baseline methods: 1) **CoT** (Wei et al. 2022): A method that instructs LLMs to generate intermediate reasoning steps (i.e., *Let's Think Step by Step*) before producing the final answer. 2) **LEAP** (Zhang et al. 2024): This method first generates incorrect reasoning samples using zero-shot CoT responses, then derives explicit guiding principles from them alongside the correct answers. During inference, the model leverages both the learned principles and provided examples to improve answer accuracy. 3) **RICP** (Sun et al. 2024): This method employs a teacher-student framework, where the teacher model identifies and analyzes the student's mistakes. During inference, relevant error cases are dynamically retrieved to construct question-specific principles for targeted guidance.

Implementation Details We construct the error database using open-source models: Qwen2.5-VL-72B-Instruct (QwenTeam 2025), DeepSeek-VL-2 (Wu et al. 2024), InternVL-3-78B (Chen et al. 2024). All experiments are conducted using four NVIDIA A100 GPUs. For closed-source models: GPT-4o (OpenAI 2024), Gemini-2.0-flash (DeepMind 2024), Claude-4-sonnet (Anthropic 2025), we set the temperature parameter to 0.7.

Main Results

As shown in Table 2, CLER consistently outperforms all baselines across all datasets. Notably, CLER's substantial improvement over the CoT baseline highlights the effectiveness of fine-grained contrastive learning combined with error reflection. On FinMR (Expertise), CLER achieves 65.90% accuracy with GPT-4o, representing a notable 8.83% gain over CoT. Both LEAP and RICP generally outperform the CoT baseline, confirming their effectiveness in improving model performance through explicit error analysis and retrieved contextual principles. Nevertheless, CLER consistently outperforms these methods, demonstrating superior capability in detailed error reflection and robustness.

When applying LEAP, we also observe several performance drops (indicated with underlines), such as GPT-4o on FinMR Math subset (-6.43%), Gemini-2.0-flash on FinMR

[1]In the broader domain of Finance, topics such as Economics and Accounting are two closely related subfields that underpin advanced financial reasoning tasks.

Model	Method	FinMR(Acc%)			FAMMA(Acc%)			MMMU-Finance(Acc%)
		ALL	Math(37.8%)	Expertise(62.2%)	ALL	Math(50%)	Expertise(50%)	Math
GPT-4o	CoT	54.17	49.40	57.07	54.50	53.00	56.00	62.77
	LEAP	57.66 (↑3.49)	42.97 (↓6.43)	**66.59 (↑9.52)**	54.50 (-)	49.00 (↓4.00)	60.00 (↑4.00)	63.56 (↑0.79)
	RICP	58.12 (↑3.95)	51.00 (↑1.60)	62.44 (↑5.37)	58.00 (↑3.50)	55.00 (↑2.00)	61.00 (↑5.00)	63.99 (↑1.22)
	CLER(ours)	**61.18 (↑7.01)**	**53.41 (↑4.01)**	65.90 (↑8.83)	**60.00 (↑5.50)**	**56.00 (↑3.00)**	**64.00 (↑8.00)**	**66.78 (↑4.01)**
Gemini-2.0-flash	CoT	56.15	50.20	59.76	56.50	53.00	60.00	56.85
	LEAP	57.81 (↑1.66)	49.80 (↓0.40)	62.68 (↑2.92)	61.00 (↑4.50)	58.00 (↑5.00)	64.00 (↑4.00)	58.03 (↑1.18)
	RICP	59.64 (↑3.49)	52.61 (↑2.41)	63.90 (↑4.14)	62.50 (↑6.00)	60.00 (↑7.00)	65.00 (↑5.00)	59.77 (↑2.92)
	CLER(ours)	**62.12 (↑5.97)**	**57.03 (↑6.83)**	**65.21 (↑5.45)**	**65.00 (↑8.50)**	**64.00 (↑11.00)**	**66.00 (↑6.00)**	**60.66 (↑3.81)**
Claude-4-sonnet	CoT	57.36	48.59	62.28	60.00	59.00	61.00	63.78
	LEAP	60.89 (↑3.53)	52.21 (↑3.62)	66.10 (↑3.82)	61.00 (↑1.00)	57.00 (↓2.00)	65.00 (↑4.00)	65.72 (↑1.94)
	RICP	61.91 (↑4.55)	53.41 (↑4.82)	67.07 (↑4.79)	64.00 (↑4.00)	62.00 (↑3.00)	66.00 (↑5.00)	66.66 (↑2.88)
	CLER(ours)	**65.10 (↑7.74)**	**56.63 (↑8.04)**	**70.24 (↑7.96)**	**67.00 (↑7.00)**	**65.00 (↑6.00)**	**69.00 (↑8.00)**	**68.77 (↑4.99)**

Table 2. Result comparison of different models and methods across all datasets. MMMU-Finance consists of math reasoning questions. ↑ for increase, ↓ for decrease. Bold indicates the best result in each column. Underlined values indicate a drop.

Math (-0.40%), and Claude-4-sonnet on FAMMA Math (-2.00%). These drops typically occur when retrieved error cases lack structural similarity to the current reasoning step, resulting in misleading principles. Such issues are common in baseline methods, underscoring the limitations of surface-level textual similarity when it is not incorporated with deeper logical structures. In contrast, CLER captures structural-semantic similarity, effectively mitigating misleading retrieval.

The MMMU-Finance dataset, which is primarily composed of math-based reasoning tasks, exhibits a slightly higher baseline accuracy. Nevertheless, CLER achieves meaningful gains, such as a 4.01% improvement over CoT with GPT-4o, showcasing the adaptability and robustness of our approach even in homogeneous, math-intensive reasoning settings. On datasets requiring substantial financial expertise (e.g., FinMR and FAMMA's expertise categories), CLER demonstrates particularly strong gains, underscoring the benefits of incorporating a detailed, fine-grained error reflection in complex multimodal reasoning tasks.

Claude-4-sonnet outperforms other models, indicating strong intrinsic reasoning capabilities in financial multimodal tasks. Despite this intrinsic strength, CLER further boosts performance, demonstrating the universal applicability and effectiveness of our method across diverse model architectures and reasoning capabilities.

Ablation Study

To assess the contribution of each module in CLER, we conduct an ablation study across all financial reasoning datasets using GPT-4o. 1) **w/o CL Retriever & Reflection**: the CoT baseline without retrieval or reflection; 2)**w/o Reflection**: step-level retrieval is performed, but MLLMs do not reflect on the retrieved information; 3) **w/o CL Retriever**: reflection is retained, but the retriever is replaced by a semantic similarity-based one; 4) **CLER**: the full method combining contrastive retrieval with step-level reflection.

Table 3 reports the performance of four configurations across all datasets. We observe consistent performance gains as more components of CLER are activated. Specifically, FinMR-All improves from 54.17% to 61.18%, while FinMR-Math increases from 49.40% to 53.41%. Similar trends are observed on FAMMA and MMMU-Finance, with gains up to +3.78% over the baseline on math-heavy subsets (e.g., 66.78% on MMMU-Math vs. 62.77%).

These results reveal three key insights: 1) Reflection is essential, as removing it results in significant drops across all subsets (e.g., FinMR-Expertise: 65.90% → 60.33%), 2) The contrastive retriever consistently outperforms semantic-based retrieval, underscoring the value of structural alignment in retrieving relevant error patterns, and 3) Gains are particularly evident in math-heavy subsets, such as FinMR-Math (49.40% → 53.41%) and MMMU-Math (62.77% → 66.78%), indicating that CLER is especially effective in correcting step-level numerical reasoning errors. These tasks often involve formula misapplication, incorrect intermediate steps, or numerical condition confusion – error types that benefit most from fine-grained structural retrieval and targeted insights. Overall, this ablation confirms that both modules—contrastive retrieval and reflection—are indispensable and synergistic within CLER.

Further Analysis and Discussion

We now delve deeper into our framework, CLER, to better understand ***why*** it achieves its improvements.

Individualized Insights of Reflection

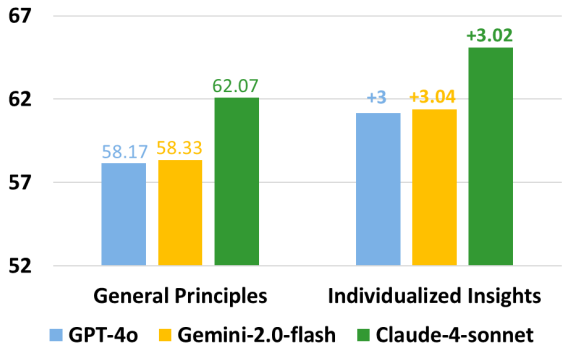

Figure 3: Performance Difference between General Principles and Individualized Insights on FinMR.

In the final stage of CLER, although error-correction pairs are retrieved from *FinErrorSet*, the model does not directly

Method	FinMR-All	FinMR-Math	FinMR-Exp	FAMMA-All	FAMMA-Math	FAMMA-Exp	MMMU-Math
w/o CL + Ref	54.17	49.40	57.07	54.50	53.00	56.00	62.77
w/o Reflection	56.78 (+2.61)	50.93 (+1.53)	60.33 (+3.26)	55.50 (+1.00)	54.00 (+1.00)	57.00 (+1.00)	62.99 (+0.22)
w/o CL Retriever	58.52 (+4.35)	51.03 (+1.63)	63.07 (+6.00)	58.00 (+3.50)	54.00 (+1.00)	62.00 (+6.00)	63.00 (+0.23)
CLER	**61.18 (+7.01)**	**53.41 (+4.01)**	**65.90 (+8.83)**	**60.00 (+5.50)**	**56.00 (+3.00)**	**64.00 (+8.00)**	**66.78 (+4.01)**

Table 3. Ablation study results on GPT-4o across FinMR, FAMMA, and MMMU-Finance.

Dataset	v3.7	v4.0	Gain
FinMR	62.85	65.10	+2.25
FinMR-Math	54.21	56.63	+2.42
FinMR-Expertise	68.10	70.24	+2.14
FAMMA	65.50	67.00	+1.50
FAMMA-Math	64.00	65.00	+1.00
FAMMA-Expertise	67.00	69.00	+2.00
MMMU-Finance	67.72	68.77	+1.05

Table 4. Improvements from Claude-3.7-sonnet to Claude-4.0-sonnet across financial datasets.

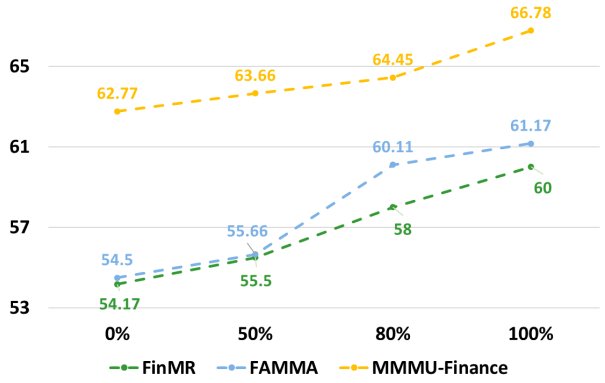

Figure 4: Effect of error database size on accuracy across FinMR, FAMMA, and MMMU-Finance.

adopt the principles generated by others. Instead, our framework encourages models to reflect *independently* on each retrieved pair. By comparing each retrieved erroneous and corrected step to its current reasoning, the model generates insights tailored to its own reasoning trajectory. This design enables the model to generate context-sensitive corrections aligned with its own reasoning structure.

To assess the impact of this design, we compare general principles (pre-generated by open-source models before testing) with individualized insights (generated by closed-source models dynamically at test time). As shown in Figure 3, tailored insights consistently improve performance across all models, with gains of 3.0% for GPT-4o, 3.04% for Gemini-2.0-flash, and 3.02% for Claude-4-sonnet. This confirms that *individualized reflection is more effective than merely borrowing fixed principles from others*.

Impact of Using Different Versions of MLLMs

To examine the effect of model version, we compare Claude 3.7-sonnet and Claude 4.0-sonnet, two successive releases from the same model family. As shown in Table 4, Claude 4.0 consistently outperforms Claude 3.7 across all evaluated benchmarks. Specifically, on FinMR overall, Claude 4.0 achieves 65.10% accuracy compared to 62.85% for Claude 3.7, reflecting a 2.25% absolute improvement. Performance gains are also evident in both FinMR-Math (2.42%) and FinMR-Expertise (2.14%), indicating improved capability in numerical and domain-specific reasoning. The improvements are consistent and robust, confirming that continual advancements in model training can synergize effectively with correction frameworks such as CLER.

Benefit of Error Database Size

To evaluate the impact of error database size on retrieval-augmented reasoning, we vary the size of the error database used in the CLER framework. Specifically, we construct three versions of the database containing 100%, 80%, and 50% of the original error instances. In each setting, we randomly subsample the database while preserving the original error-type distribution to ensure fairness. As shown in Figure 4, reducing the size of the error database results in a clear and consistent decline in performance across all benchmarks. These results show that database scale is critical for effective retrieval and correction. A larger database not only increases the likelihood of retrieving highly relevant errors but also offers more diverse and informative counterexamples, thereby enhancing model reflection and learning.

Conclusion

We introduce CLER, a novel framework that enhances multimodal financial reasoning by reflecting cross-model errors. CLER retrieves step-level mistake–correction pairs from a curated error database and prompts model-specific reflection, enabling robust and context-aware reasoning. Additionally, CLER generates the *FinErrorSet* used by free open-source models, resulting in a lower cost compared to using commercial MLLMs to prepare an error dataset. Experiments on financial benchmarks show that CLER consistently outperforms strong baselines in both math and expertise reasoning tasks, with ablation studies confirming the complementary roles of contrastive retrieval and structured reflection. **Limitations:** CLER's performance depends on the quality and diversity of the error database, which may limit generalization to novel error types. It also focuses on step-level reflection, potentially overlooking cross-step or higher-order reasoning patterns.

References

An, S.; Ma, Z.; Cai, S.; Lin, Z.; Zheng, N.; Lou, J.-G.; and Chen, W. 2024. Can LLMs Learn From Mistakes? An Empirical Study on Reasoning Tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 833–854.

Anthropic. 2025. Claude Sonnet 4.

Bandura, A. 1977. *Social Learning Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bhatia, G.; Nagoudi, E. M. B.; Cavusoglu, H.; and Abdul-Mageed, M. 2024. FinTral: A Family of GPT-4 Level Multimodal Financial Large Language Models. arXiv:2402.10986.

Chen, Z.; Chen, W.; Smiley, C.; Shah, S.; Borova, I.; Langdon, D.; Moussa, R.; Beane, M.; Huang, T.-H.; Routledge, B.; and Wang, W. Y. 2021. FinQA: A Dataset of Numerical Reasoning over Financial Data. *Proceedings of EMNLP 2021*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*.

Chia, Y. K.; Chen, G.; Tuan, L. A.; Poria, S.; and Bing, L. 2023. Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

DeepMind, G. 2024. Gemini 2.0 Flash Model Card.

Deng, S.; Peng, H.; Xu, J.; Liu, C.; Giurcaneanu, C. D.; and Liu, J. 2025. Understanding Financial Reasoning in AI: A Multimodal Benchmark and Error Learning Approach. arXiv:2506.06282.

Du, K.; Mao, R.; Xing, F.; and Cambria, E. 2024. Explainable Stock Price Movement Prediction using Contrastive Learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, 529–537. Idaho, USA.

Du, K.; Zhao, Y.; Mao, R.; Xing, F.; and Cambria, E. 2025a. Natural Language Processing in Finance: A Survey. *Information Fusion*, 115: 102755.

Du, K.; Zhao, Y.; Mao, R.; Xing, F.; and Cambria, E. 2025b. A Retrieval-Augmented Multi-Agent System for Financial Sentiment Analysis. *IEEE Intelligent Systems*, 40: 15–22.

Durkin, K.; and Rittle-Johnson, B. 2012. Learning from errors: Differences in how students reason about and learn from incorrect examples. *Journal of Educational Psychology*, 104(2): 289–305.

Gan, Z.; Lu, Y.; Zhang, D.; Li, H.; Liu, C.; Liu, J.; Liu, J.; Wu, H.; Fu, C.; Xu, Z.; Zhang, R.; and Dai, Y. 2024. MME-Finance: A Multimodal Finance Benchmark for Expert-level Understanding and Reasoning. arXiv:2411.03314.

Guo, Z.; Xu, R.; Yao, Y.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.-S.; Liu, Z.; and Huang, G. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, 390–406. Springer.

Jin, L.; Lin, B.; Hong, M.; Zhang, K.; and So, H.-J. 2025. Exploring the impact of an llm-powered teachable agent on learning gains and cognitive load in music education. *arXiv preprint arXiv:2504.00636*.

Li, Y.; Yuan, P.; Feng, S.; Pan, B.; Sun, B.; Wang, X.; Wang, H.; and Li, K. 2024a. Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18591–18599.

Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; and Bai, X. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26763–26773.

Liu, C.; Arulappan, A.; Naha, R.; Mahanti, A.; Kamruzzaman, J.; and Ra, I.-H. 2024. Large Language Models and Sentiment Analysis in Financial Markets: A Review, Datasets, and Case Study. *IEEE Access*, 12: 134041–134061.

Luo, J.; Kou, Z.; Yang, L.; Luo, X.; Huang, J.; Xiao, Z.; Peng, J.; Liu, C.; Ji, J.; Liu, X.; Han, S.; Zhang, M.; and Guo, Y. 2025. FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Mao, R.; He, K.; Ong, C. B.; Liu, Q.; and Cambria, E. 2024. MetaPro 2.0: Computational Metaphor Processing on the Effectiveness of Anomalous Language Modeling. In *Findings of the Association for Computational Linguistics: ACL*. Bangkok, Thailand.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

QwenTeam. 2025. Qwen2.5-VL.

Ranga, S.; Mao, R.; Bhattacharjee, D.; Cambria, E.; and Chattopadhyay, A. 2024. RTL Agent: An Agent-Based Approach for Functionally Correct HDL Generation via LLMs. In *IEEE 33rd Asian Test Symposium (ATS)*. Ahmedabad, Gujarat, India.

Rangapur, A.; Wang, H.; Jian, L.; and Shu, K. 2025a. FinFact: A Benchmark Dataset for Multimodal Financial Fact-Checking and Explanation Generation. In *Companion Proceedings of the ACM on Web Conference 2025*, 785–788.

Rangapur, A.; Wang, H.; Jian, L.; and Shu, K. 2025b. FinFact: A Benchmark Dataset for Multimodal Financial Fact-Checking and Explanation Generation. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, 785–788. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713316.

Sawhney, R.; Mathur, P.; Mangal, A.; Khanna, P.; Shah, R. R.; and Zimmermann, R. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*, 456–465.

Sun, H.; Jiang, Y.; Wang, B.; Hou, Y.; Zhang, Y.; Xie, P.; and Huang, F. 2024. Retrieved in-context principles from previous mistakes. *arXiv preprint arXiv:2407.05682*.

Tong, Y.; Li, D.; Wang, S.; Wang, Y.; Teng, F.; and Shang, J. 2024. Can LLMs learn from previous mistakes? Investigating LLMs' errors to boost for reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3065–3080. Bangkok, Thailand: Association for Computational Linguistics.

Tulis, M. 2013. Error-related learning and the negative knowledge effect. *Psychological Research*, 77(6): 752–762.

Van Gog, T.; and Rummel, N. 2010. Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*, 22(2): 155–174.

Wang, R.; Li, H.; Han, X.; Zhang, Y.; and Baldwin, T. 2024. Learning from failure: Integrating negative examples when fine-tuning large language models as agents. *arXiv preprint arXiv:2402.11651*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; Xie, Z.; Wu, Y.; Hu, K.; Wang, J.; Sun, Y.; Li, Y.; Piao, Y.; Guan, K.; Liu, A.; Xie, X.; You, Y.; Dong, K.; Yu, X.; Zhang, H.; Zhao, L.; Wang, Y.; and Ruan, C. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. arXiv:2412.10302.

Xie, Q.; Han, W.; Zhang, X.; Lai, Y.; Peng, M.; Lopez-Lira, A.; and Huang, J. 2023. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance. arXiv:2306.05443.

Xue, S.; Li, X.; Zhou, F.; Dai, Q.; Chu, Z.; and Mei, H. 2024. Famma: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint arXiv:2410.04526*.

Xue, S.; Li, X.; Zhou, F.; Dai, Q.; Chu, Z.; and Mei, H. 2025. FAMMA: A Benchmark for Financial Domain Multilingual Multimodal Question Answering. arXiv:2410.04526.

Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023. FinGPT: Open-Source Financial Large Language Models. arXiv:2306.06031.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9556–9567.

Zhang, T.; Madaan, A.; Gao, L.; Zheng, S.; Mishra, S.; Yang, Y.; Tandon, N.; and Alon, U. 2024. In-context principle learning from mistakes. *arXiv preprint arXiv:2402.05403*.

Zhang, X.; and Yang, Q. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 4435–4439.

Zhao, H.; Liu, Z.; Wu, Z.; Li, Y.; Yang, T.; Shu, P.; Xu, S.; Dai, H.; Zhao, L.; Mai, G.; et al. 2024a. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

Zhao, S.; Wang, Z.; Juefei-Xu, F.; Xia, X.; Liu, M.; Wang, X.; Liang, M.; Zhang, N.; Metaxas, D. N.; and Yu, L. 2025. Accelerating Multimodal Large Language Models by Searching Optimal Vision Token Reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29869–29879.

Zhao, Y.; Liu, H.; Long, Y.; Zhang, R.; Zhao, C.; and Cohan, A. 2024b. FinanceMath: Knowledge-Intensive Math Reasoning in Finance Domains. arXiv:2311.09797.

Zhao, Y.; Liu, H.; Long, Y.; Zhang, R.; Zhao, C.; and Cohan, A. 2024c. FinanceMATH: Knowledge-Intensive Math Reasoning in Finance Domains. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12841–12858. Bangkok, Thailand: Association for Computational Linguistics.

Zhao, Y.; Long, Y.; Liu, H.; Kamoi, R.; Nan, L.; Chen, L.; Liu, Y.; Tang, X.; Zhang, R.; and Cohan, A. 2023. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. *arXiv preprint arXiv:2311.09805*.

Zhou, Z.; Tao, R.; Zhu, J.; Luo, Y.; Wang, Z.; and Han, B. 2024. Can Language Models Perform Robust Reasoning in Chain-of-thought Prompting with Noisy Rationales? *Advances in Neural Information Processing Systems*, 37: 123846–123910.

Zhu, F.; Lei, W.; Huang, Y.; Wang, C.; Zhang, S.; Lv, J.; Feng, F.; and Chua, T.-S. 2021. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3277–3287. Online: Association for Computational Linguistics.