PyFi: Toward Pyramid-like Financial Image Understanding for VLMs via Adversarial Agents

Anonymous ACL submission

Abstract

This paper proposes *PyFi*, a novel framework for pyramid-like financial image understanding that enables vision language models (VLMs) to reason through question chains in a progressive, simple-to-complex manner. At the core of PyFi is ***PyFi*-600K**, a dataset comprising 600K financial question-answer pairs organized into a reasoning pyramid: questions at the base require only basic perception, while those toward the apex demand increasing levels of capability in financial visual understanding and expertise. This data is scalable because it is synthesized without human annotations, using *PyFi*-adv, a multi-agent adversarial mechanism under the Monte Carlo Tree Search (MCTS) paradigm, in which, for each image, a challenger agent competes with a solver agent by generating question chains that progressively probe deeper capability levels in financial visual reasoning. Leveraging this dataset, we present fine-grained, hierarchical, and comprehensive evaluations of advanced VLMs in the financial domain. Moreover, fine-tuning Qwen2.5-VL-3B and Qwen2.5-VL-7B on the pyramid-structured question chains enables these models to answer complex financial questions by decomposing them into sub-questions with gradually increasing reasoning demands, yielding average accuracy improvements of 19.52% and 8.06%, respectively, on the dataset.

1 Introduction

Vision language models (VLMs) are capable of step-by-step reasoning, presented as Chain-of-Thought (CoT) (Wei et al., 2022), for image understanding in question answering (Q&A) across various scientific domains (Chen et al., 2024a; Yue et al., 2024; Wu et al., 2025d), driven by the development of domain-specific benchmarks (Weng et al., 2025; Wang et al., 2025b; Yue et al., 2024; Wang et al., 2025a), step-wise annotations (Wang et al., 2024c; Zhang et al., 2024b; Qi et al., 2024), and domain-tailored VLMs (Yang et al., 2023; Huang et al., 2024; Lan et al., 2025). In contrast, advancements in the financial domain remain limited. This is because, in finance image understanding, the creation of CoT reasoning datasets, which are highly necessary for the evaluation and fine-tuning of VLMs, requires a high level of expertise and domain knowledge and is therefore scarce.

To address these issues, previous literature (Chen et al., 2021) collects Q&A samples from financial documents manually, a process that is prohibitively time-consuming and difficult to scale. As a result, subsequent works propose datasets by merging extensive existing financial datasets (Shah et al., 2022; Xie et al., 2024). These repurposed datasets may introduce accumuated biases and inconsistencies inherited from the original sources. More recently, benchmarks such as (Gan et al., 2025; Kim et al., 2024; Luo et al., 2025; Lu et al., 2025; Li et al., 2025a; Liu et al., 2025d) leverage trained VLMs to generate Q&A pairs augmented with CoTs (Xue et al., 2025). However, these samples may contain factual errors or hallucinations, which can distort VLM evaluation and fine-tuning and thereby undermine their reliability in real-world financial applications, where precision and expert-level understanding are essential.

Therefore, this paper proposes a framework named *PyFi*, which supports comprehensive evaluation and advanced fine-tuning of VLMs through a pyramid-like hierarchy for financial image understanding. Specifically, *PyFi* includes a synthesized dataset, ***PyFi-600K*** shown in Figure 1, comprising 600 thousand (K) samples with step-wise annotations that provide expert-level guidance on how to progressively solve complex financial problems across 17 categories. The construction of this dataset is enabled by *PyFi*-adv, a multi-agent adversarial mechanism designed to automatically synthesize and refine samples using expert-level financial knowledge.

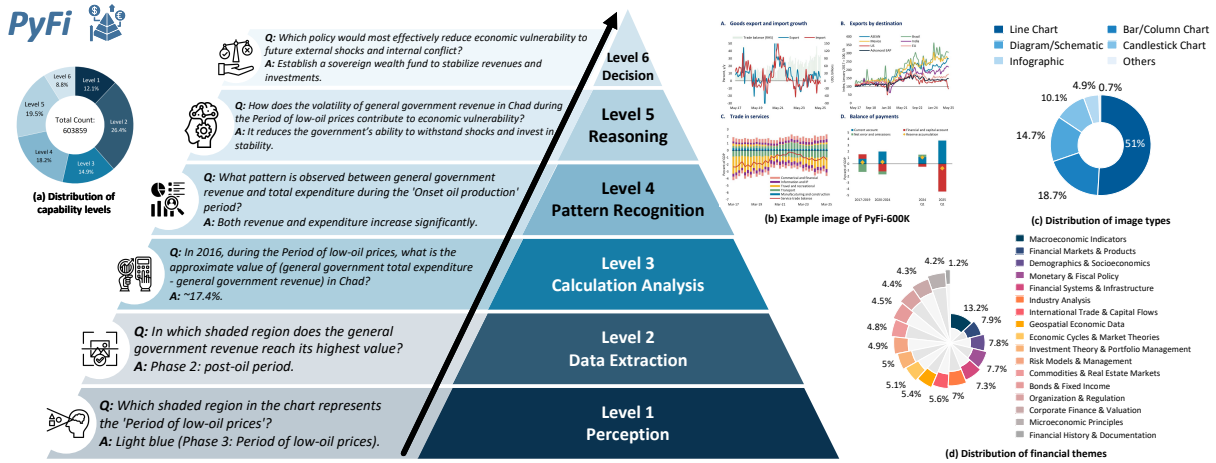


Figure 1: Overview of the ***PyFi*-600K** dataset within our *PyFi* framework. The dataset is structured as a pyramid comprising 6 capability levels, 11 financial image types, and 17 financial themes.

PyFi-600K offers three key benefits for the interpretability and fine-tuning of VLMs in finance. First, the samples are organized in a pyramid structure, where the required image understanding capability increases progressively — from basic visual perception at the bottom level to complex financial decision-making at the top. This hierarchy, with capability levels ranging from one to six, enables comprehensive evaluation of VLMs across multiple aspects and supports ranking the capabilities required by VLMs in different financial themes. Second, samples across different levels are interconnected through *question chains*, demonstrating how questions at higher levels can be addressed step-by-step by leveraging information and reasoning from lower-level Q&A samples. Finally, it supports process supervision (Lightman et al., 2023), as each sample is assigned a reward score that reflects the correctness of its reasoning process in solving the financial problem.

The construction of such a scalable dataset is enabled by our proposed *PyFi*-adv, a multi-agent adversarial mechanism in which a challenger agent progressively synthesizes increasingly difficult questions, while a solver agent generates answers by reasoning with financial knowledge. The two agents compete within a Monte Carlo Tree Search (MCTS) framework: the challenger begins with simple questions at level 1, either selecting existing ones or generating new ones, and presents them to the solver for response. Based on the answer, the challenger then formulates a more challenging question in the next round. This iterative process continues until the information accumulated suffices to address the decision-making task at level 6.

Our extensive experiments with ***PyFi-600K*** reveal that 15 well-known VLMs struggle to answer complex financial questions, with average accuracy dropping from 71.80% at level 1 to 32.95% at level 6. With *question chains*, we observe that addressing a Level-6 financial decision-making question typically requires correctly answering an average of 10.48 sub-questions from lower levels. In particular, our results highlight that errors in calculation analysis are a primary cause of VLMs' failure in final financial decision-making. More importantly, fine-tuning Qwen-VL models on ***PyFi-600K*** using *question chains* improves accuracy by an average of 13.79%, with gains reaching up to 19.52% in the best case.

2 Related Work

Financial Benchmark. The evaluation of Vision-Language Models (VLMs) in the finance domain has seen remarkable progress, from early text-based benchmarks (Chen et al., 2021; Shah et al., 2022; Xie et al., 2024) to more recent efforts targeting financial image understanding (Kim et al., 2024; Zhu et al., 2025a; Shu et al., 2025; Li et al., 2025a; Liu et al., 2025d; Rangapur et al., 2025; Zhao et al., 2025; Gan et al., 2025; Luo et al., 2025; Xue et al., 2025; DENG et al., 2025). However, the question-answering datasets of most benchmarks, such as FinMME (Gan et al., 2025), rely on human annotation, making them costly and difficult to scale. Besides, they fail to support comprehensive assessment of VLMs in the financial domain. However, we introduce an automatically synthesized dataset for the fine-grained, hierarchical evaluation

of financial visual understanding across diverse image types and financial themes. More related background can be found in Appendix A.

Financial Large Models. Previous efforts primarily centered on text analysis using financial LMs (Huang et al., 2023; Wu et al., 2023; Yang et al., 2023; Xie et al., 2023; Liu et al., 2025c). Recent financial VLMs (Wang et al., 2023; Bhatia et al., 2024; Huang et al., 2024; Park et al., 2025) integrate visual inputs to facilitate step-by-step reasoning, thereby enabling broader financial decision-making. However, these models lack the expertise and interpretability needed for challenging financial themes, a limitation stemming from the absence of detailed, step-wise problem-solving processes grounded in finance-specific knowledge. In contrast, our dataset contains questions that illustrate how to address financial problems through a progression from simple to complex image understanding and from basic to increasingly deep financial knowledge.

Multi-Agent LM Framework. Recent advances in LMs have enabled Agentic AI to excel in real-world applications, with multiple LM-based agents now collaborating to perceive, learn, reason, and act, supporting intricate tasks such as simulation (Lin et al., 2025; ShengbinYue et al., 2025), reasoning (Liu et al., 2024a; Wu et al., 2025a; Du et al., 2024), trading (Li et al., 2025b; Yu et al., 2024; Wu et al., 2025c), and data synthesis (Tang et al., 2024; Liu et al., 2025a; Xuan et al., 2025). However, to our knowledge, no existing method in finance leverages a multi-agent adversarial approach to synthesize data. In contrast, inspired by Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), we present the first method that employs adversarial agents to automatically synthesize hierarchical, fine-grained financial datasets.

3 Framework

3.1 Financial Dataset in Pyramid Structure

We release ***PyFi-600K***, a large-scale finance dataset comprising 600 thousand (K) samples, each consisting of a question-answer pair, constructed within a pyramid structure that encompasses six levels of finance-specific visual cognition: *Perception* (*PP*), *Data Extraction* (*DE*), *Calculation Analysis* (*CA*), *Pattern Recognition* (*PR*), *Logical Reasoning* (*LR*), *Decision Support* (*DS*). More importantly, ***PyFi-600K*** introduces *sample chains* that demonstrate how to address higher-level pyramid questions by leveraging knowledge from lower-level samples.

Specifically, we denote each finance sample as $\boldsymbol{s}^l = (I, Q, A, r)$, where $r \in [0, 1]$ is the reward reflecting its correctness, I is the finance image to understand, $l \in [1, 6]$ indicates the pyramid level to which the question Q and the answer A belong. The sample complexity degree is $C\left(\boldsymbol{s}^l\right) \in [1, 5]$. And, as presented in right sub-figure of Figure 2, we define *sample chain* as $\boldsymbol{S}^l_{1\ldots n} = \left[\boldsymbol{s}^1_1, ..., \boldsymbol{s}^l_n\right]$ where n is the length and $n \geq l$ because samples within the chain can come from the same level. Considering the chain, the sequence of questions increases in levels or complexity degrees, as the answer for each one requires more insightful image understanding and financial expertise than the preceding ones. That is, for any $\boldsymbol{s}^h_i$ and $\boldsymbol{s}^k_j$ with $i, j \in [1, \ldots, n]$, $h, k \in [1, \ldots, l]$, $i < j$, we have that $h \leq k$, and if $h = k$, then $C\left(\boldsymbol{s}^h_i\right) \leq C\left(\boldsymbol{s}^k_j\right)$. Subsequently, the reasoning process for answering the question in any sample within the chain depends on its predecessor samples. Specifically, when $\boldsymbol{s}^l_j$ is addressed by a VLM parameterized by $\boldsymbol{\theta}$, we have $p_{\boldsymbol{\theta}}\left(A|I, Q\right) < p_{\boldsymbol{\theta}}\left(A|I, Q, \left\{\boldsymbol{s}^l_i\right\}^j_{i=1}\right)$, meaning that the reasoning information and financial knowledge in answers from predecessor samples enable more reliable question answering for the current sample.

Therefore, compared to existing manual and automatic evaluation dataset construction methods, we observe the following three advantages of the proposed approach for the finance domain:

Hierarchical Cognition in the dataset embodies human-level problem solving processes, progressing from basic perception to complex financial decision-making. Samples at each level l correspond to a specific cognitive capability, which builds upon those from lower levels $i = 1.., l-1$ or complexities. Thus, evaluating VLMs on these hierarchically structured samples enables a comprehensive assessment of their abilities in the finance domain. That is, any financial reasoning performed for questions from higher levels, which require rare, expert-level knowledge, relies on foundational financial questions involving base cognitions such as *CA*. Besides, we can identify how VLMs perform, whether better or worse, at each cognition level during financial reasoning, thereby revealing which level is crucial for which financial questions and to what extent.

Progressive Logical Chain in the dataset is pre-

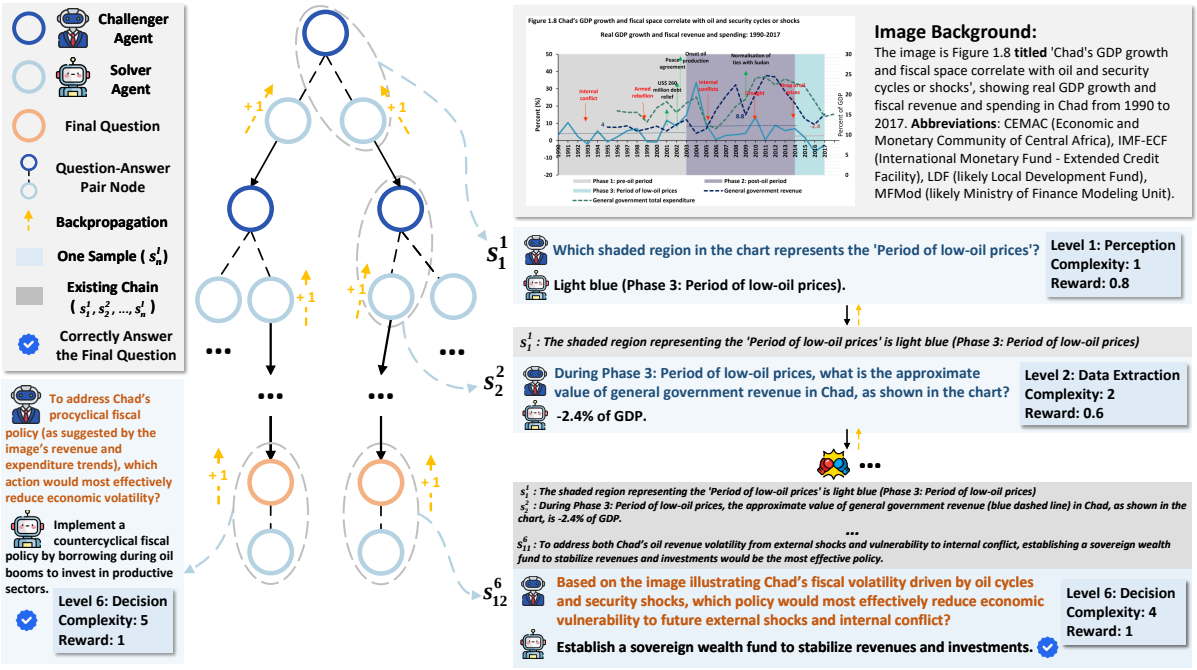


Figure 2: Overview of *PyFi*-adv: a challenger agent competes with a solver agent under the Monte Carlo tree search (MCTS) paradigm to generate question chains that progressively probe deeper capability levels in financial visual reasoning. See the complete chain in Figure 12 in Appendix.

sented as the *sample chain* $\boldsymbol{S}_{1\ldots n}^{l}$ shown in Figure 2, which demonstrates how a challenging financial question at level $l = 6$ is solved by progressively addressing a sequence of simple-to-complex sub-questions. This aligns with financial decision-making logic, where each step builds upon the previous one to form a cumulative and logically coherent reasoning flow. Such a traceable chain enhances the interpretability and justifiability required in the finance domain. Moreover, VLMs fine-tuned on such chains are capable of solving complex financial problems by accumulating knowledge through first addressing a series of simpler problems.

Step-wise Reward Score r of each sample reflects its reliability; thus, the scores along a sample chain support process supervision (Lightman et al., 2023) in reasoning with financial image understanding. Using these scores, we enable the training of process reward models that act as verifiers for any financial VLM.

3.2 Automatic Synthesization with Adversarial Agents

Existing financial datasets rely solely on human annotation, while automated methods from other domains, such as rStar-Math (Guan et al., 2025) for mathematical reasoning, fail to meet the requirements of finance, where interpretability and reliable reasoning are essential, along with increasing complexity, domain expertise, and hierarchical cognition. We achieve this by proposing a multi-agent adversarial mechanism, referred to as *PyFi*-adv, which constructs question chains, as illustrated in Figure 2.

Specifically, *PyFi*-adv contains a challenger agent that is supported by an off-and-shelf VLM, denoted as ψ, and similarly, a solver agent ϕ. Inspired by the adversarial mechanism in AlphaGo (Silver et al., 2016), our two agents compete within an MCTS framework to automatically synthesize high-quality reasoning samples. In the finance domain, this approach substantially outperforms existing automatic data augmentation methods discussed in subsection A.3.

Confrontation for Sample Synthesization, as shown by the Gray oval circle in Figure 2, is that when ϕ aims to correctly answer the given question, ψ checks the answer to propose a more challenging follow-up question to hinder the solver. This is formulated as $Q_i \sim p_\psi\left(Q_i | I, \mathbf{P}_q, \boldsymbol{S}_{1\ldots i-1}^{l}\right)$ and thus $A_i \sim p_\phi(A_i | I, Q_i, \mathbf{P}_a, \boldsymbol{S}_{1\ldots i-1}^{l})$, where $\mathbf{P}_q$ and $\mathbf{P}_a$ are the agent prompts.

Exploration for Diversity Guarantee aims to prevent the adversarial process from converging to a set of questions with identical content. At each step, a Bernoulli sampling is used to determine the ψ whether to *explore* a new question path or *exploit* an existing one. For exploitation, the UCT

Table 1: Evaluation of Pre-trained VLMs on ***PyFi*-600K** across 6 capability levels and 5 complexity degrees, conducted on the **Evaluation Dataset after Exclusion**. This dataset is derived from an initial random selection of 1000 instances from PyFi-600K, followed by an additional filtering step to **address potential data leakage**. The final test set consists of 301 samples. For simplicity, we only present the results from complexity degrees 3, 4, and 5. The full model names listed in the "Model" column are: GPT-4.1, InternVL3-38B, Claude-opus-4-1-20250805, DeepSeek-VL2, Hunyuan-Large-Vision, ERNIE-4.5-turbo-vl, Moonshot-V1-128k-Vision-Preview, Moonshot-V1-32k-Vision-Preview, Moonshot-V1-8k-Vision-Preview, Qwen3-VL-Plus, Qwen2.5-VL-72B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen2.5-VL-7B-Instruct, and Qwen2.5-VL-3B-Instruct. Additionally, we also evaluated recent state-of-the-art models, including GPT-5.5, Claude-opus-4.7, Gemini-3.5-flash, and Kimi-k2.5.

Model	**Overall**	***PP***	***DE***	***CA***	***PR***	***LR***	***DS***	3	4	5
GPT-4.1 (OpenAI, 2025)	52.99	86.84	47.64	40.62	57.14	54.84	38.46	52.63	46.55	53.85
InternVL3 (Zhu et al., 2025b)	52.91	76.32	56.37	44.53	43.88	53.23	30.77	41.78	44.83	67.31
Claude-opus-4.1 (Anthropic, 2024)	64.70	80.92	62.74	**66.02**	58.16	72.58	32.69	66.12	55.60	73.08
GLM-4.5V (Hong et al., 2025)	**74.75**	89.47	**78.07**	65.23	**75.51**	**75.81**	46.15	**72.37**	**64.22**	73.08
DS-VL2 (Wu et al., 2024b)	45.18	76.32	44.58	36.33	42.35	42.74	19.23	34.54	42.24	46.15
Hunyuan-LV (Sun et al., 2024)	59.72	**90.79**	58.96	36.72	63.27	69.35	**51.92**	50.00	57.76	**80.77**
ER-4.5-TVL (Baidu-ERNIE-Team, 2025)	34.47	34.87	33.96	36.72	25.00	45.16	36.54	38.82	25.43	65.38
MV1-128k (Team et al., 2025b)	54.57	80.92	61.08	39.06	41.84	58.87	38.46	46.71	43.97	57.69
MV1-32k (Team et al., 2025b)	54.40	80.26	61.56	37.11	40.31	62.90	38.46	45.72	43.97	61.54
MV1-8k (Team et al., 2025b)	54.90	80.26	61.79	39.84	42.35	58.06	38.46	48.03	43.97	55.77
Q3-VL-Plus (Yang et al., 2025a)	51.00	77.63	54.25	37.50	41.84	57.26	32.69	45.72	37.50	65.38
Q2.5-VL-72B (Bai et al., 2025)	48.84	70.39	59.91	33.20	35.71	46.77	26.92	36.51	31.03	57.69
Q2.5-VL-32B (Bai et al., 2025)	43.19	60.53	45.28	37.50	36.73	42.74	28.85	39.14	30.60	59.62
Q2.5-VL-7B (Bai et al., 2025)	37.87	54.61	49.53	34.77	18.37	24.19	15.38	31.91	17.67	38.46
Q2.5-VL-3B (Bai et al., 2025)	20.51	36.84	15.57	21.48	16.84	21.77	19.23	23.03	18.97	30.77
GPT-5.5 (OpenAI, 2026)	70.68	86.84	73.11	66.80	63.78	73.39	42.31	65.79	63.36	71.15
Claude-opus-4.7 (Anthropic, 2026)	70.18	84.87	70.99	69.53	61.73	75.81	42.31	66.78	61.64	75.00
Gemini-3.5-flash (Google, 2026)	70.68	86.84	69.81	69.53	66.33	73.39	46.15	65.79	67.24	71.15
Kimi-k2.5 (Team et al., 2026)	54.65	80.26	57.55	42.97	51.53	52.42	30.77	47.04	43.10	69.23

formula is employed to select the most promising prior question.

Expansion for Hierarchical Questions is naturally achieved through the adversarial agents: once ψ performs a *question expansion*, it generates a subsequent, more complex question based on the accumulated $\boldsymbol{S}^l_{1\ldots i}$. Agent ϕ then produces an answer A_{i+1}, forming a new sample $\boldsymbol{s}^l_{i+1}$. Thus, a hierarchical question structure is built through repeated adversarial interactions that continue until either the top-level ($l = 6$) question is addressed or a predefined depth limit is reached.

Backpropagation for Reward Assignment adjusts the success rate of each sample by backpropagating the comparison between the answer of the top-level question and the ground truth down to individual samples in the chain, as indicated by the orange dotted line in Figure 2. Upon convergence of the MCTS, the resulting success rate is used as the reward score, providing a correctness label that measures the reliability of the synthesized dataset.

4 Experiments

4.1 Setup

Datasets. *PyFi*-600K dataset consists 603,859 samples and 62,660 sample chains with reward scores of 1. For evaluation, we proportionally sample 1,000 samples from each capability level to construct the test set across all cases. See detail in Appendix B.

Data Leakage Mitigation. To ensure a fair evaluation, we systematically identify and exclude samples that exhibit data leakage. Specifically, any question that can be correctly answered by more than half of the evaluated models without the image is considered leaked. This reduces our test set from 1,000 to *301* samples. Detailed analysis and justification are provided in Appendix D.

Models. For evaluation, we use 15 VLMs as listed in Table 1. For SFT, we fine-tune Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct on sample chains to obtain *PyFi-QwenVL-3B* and *PyFi-QwenVL-7B*, respectively.

SFT. We fine-tuned the models on approximately 47K sample chains using two dataset variants: one with chain-of-thought (CoT) annotations (w/ CoT) and one without (w/o CoT). In the CoT variant, each question chain is converted into a CoT reasoning sequence, where each reasoning step corresponds to a sub-question and its answer. Thus, a chain with n samples yields a CoT sequence with n reasoning steps. In the non-CoT variant, only the question and answer from the final sample in the chain are used. We use the AdamW optimizer with a learning rate of 1.0×10^{-4}, cosine learning rate scheduling, and a warmup ratio of 0.1. Train-

ing is conducted for one epoch with an effective batch size of 8. For parameter-efficient fine-tuning, we apply LoRA with full-module adaptation. All experiments are carried out on four NVIDIA RTX 5090 GPUs.

Evaluation. We prompt all models to place the solution within `\boxed{}`. For all inference runs, we use a temperature of 0.1 and a top-p value of 1.0. Accuracy is reported as the evaluation metric.

4.2 Main Results

Evaluations of Pre-trained VLMs are presented in Table 1. In summary, we observe that as the capability level and complexity of financial image understanding increase, the accuracy of VLMs gradually declines. Specifically, the average accuracy across all models decreases from 71.80% at level 1 (Perception) to 32.95% at level 6 (Decision Support). This indicates that while current VLMs can handle basic visual perception in the financial image understanding, they struggle with complex financial questions that require higher-level cognitive abilities. Notably, GLM-4.5V achieves the highest overall accuracy of 74.75%, demonstrating its strong performance across various capability levels. However, even this model's accuracy drops to 46.15% at level 6, highlighting the challenges in financial decision-making. Most models, including GPT-4.1 and InternVL3-38B, achieve over 70% accuracy at level 1 but fall below 40% at level 6, resulting in a relatively low average overall accuracy of 50%. Recent VLMs (GPT-5.5, Claude-opus-4.7, Gemini-3.5-flash, and Kimi-k2.5) achieve up to 70.68% overall accuracy, but like prior models, drop sharply from level 1 to level 6.

Evaluations of Fine-tuned Qwen-VL models are presented in Figure 3. In summary, with *question chains* arranged in a pyramid manner, the models, especially the smaller ones, show a significant increase in accuracy and gain the ability to reason progressively, moving from low-level to high-level questions until they reach the solution, as shown in Figure 5. Specifically, both our *PyFi-* models exhibit substantial accuracy gains. *PyFi-CoT* related models improves the 3B model by 19.52% points and the 7B model by 8.06% points over their local baselines. The gain is notably larger for the smaller 3B model. Notably, on higher-level financial questions, such as Decision Support (DS), *PyFi-CoT* ones improve the 3B model by 9.62% and the 7B model by 23.08%. The substantially larger gain for the 7B VLM suggests that

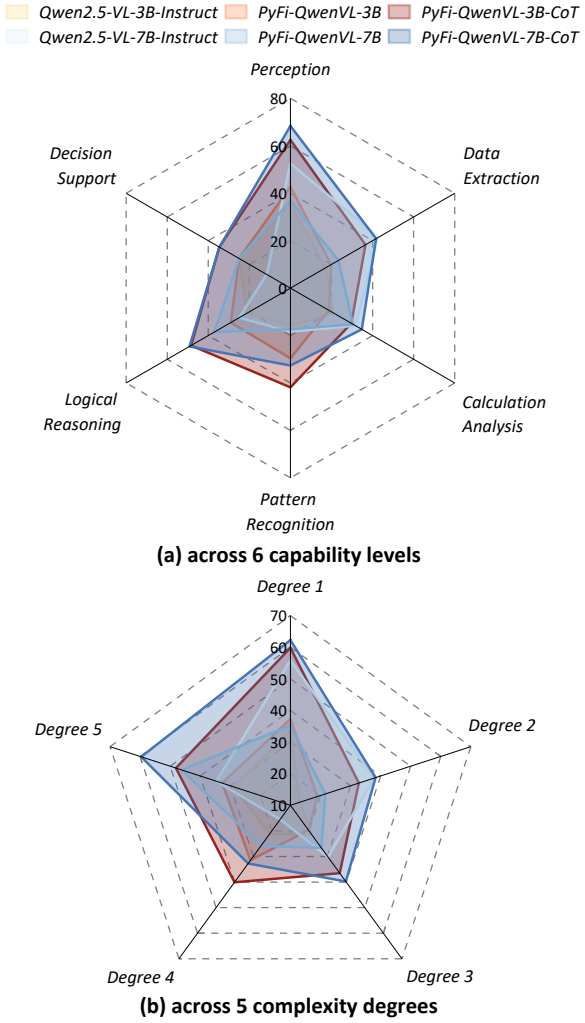


Figure 3: Comparison between Qwen2.5-VL models and ours (*PyFi-QwenVL-3B* and *PyFi-QwenVL-7B*).

larger VLMs benefit more from question-chain fine-tuning in challenging financial scenarios. Notably, in Pattern Recognition (PR) and Logical Reasoning (LR), *PyFi-QwenVL-3B* even surpasses the accuracy of the much larger 72B VLM, as shown in Table 1. Building on the detailed comparison in Figure 5, we argue that sample-chain fine-tuning equips VLMs with the ability to solve complex financial problems by iteratively addressing a sequence of self-generated sub-questions of increasing capability levels, thereby significantly improving interpretability and reliability.

4.3 Main Benefits and Insights

Evaluations in Table 1 and Figure 3 enable a fine-grained analysis of how — and to what extent — VLMs meet the varying capability requirements of financial image understanding across different levels of complexity. The phenomenon of decreasing accuracy as the level increases in Table 1 indicates

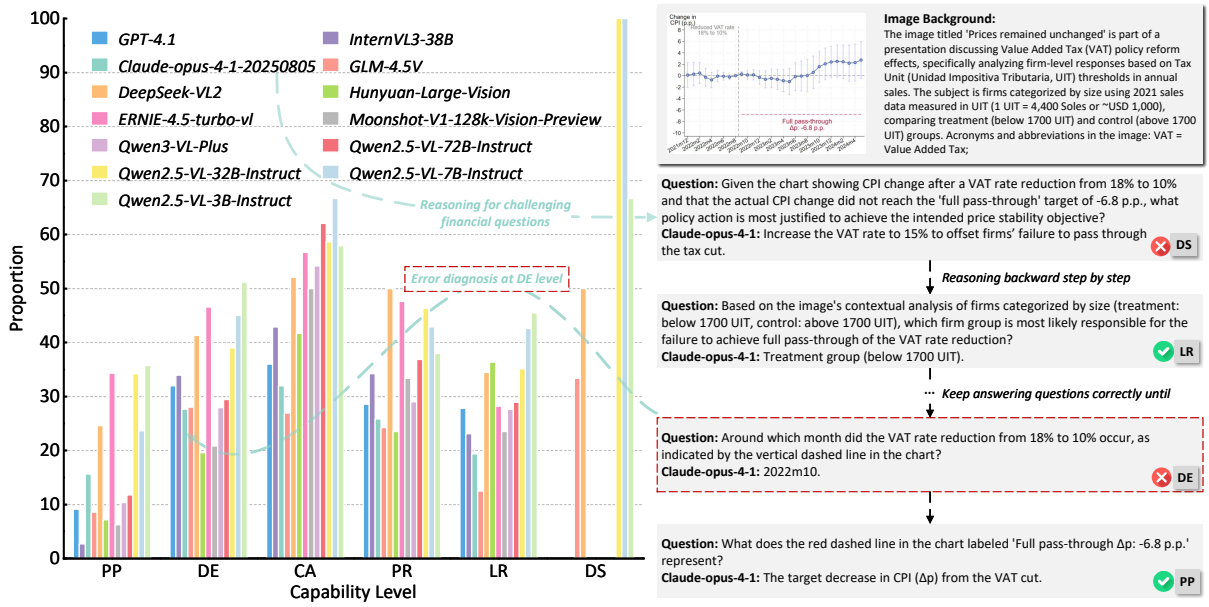


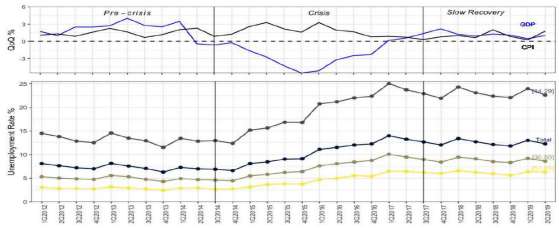
Figure 4: Evaluation of error proportions across capability levels leading to incorrect financial decisions. As illustrated in the right sub-figure, the *sample chain* enables error tracing: when Claude-Opus-4-1 fails to answer the Level 6 financial decision-making question, we can **trace back** to identify the specific lower-level sub-questions that were answered incorrectly, thereby pinpointing the exact level at which the error originated.

that existing VLMs can only meet the most basic image understanding in finance and are difficult to be used for complex financial decision-making. For instance, while models like GLM-4.5V perform well on basic perception questions (level 1), their accuracy drops considerably on decision support questions (level 6). This trend is consistent across most models, suggesting that greater efforts are needed to enhance VLMs' capabilities in the finance domain. Thus, the results in Figure 3 demonstrate the importance of our question-chain fine-tuning: Qwen models learn to make accurate financial decisions by first addressing a series of sub-questions in a progression from simple to complex, a process aligned with human cognition.

Tracing back from Decision Support to Lower-level Q&As, as illustrated in Figure 4, significantly improves the interpretability of VLMs in financial decision-making. In Figure 4, we first collect Level-6 samples that VLMs fail to answer correctly and analyze their *question chains* by evaluating VLMs on the questions at each lower level in the chain. In the finance domain, addressing a Level-6 financial decision-making question typically requires correctly answering sub-questions from lower levels. Thus, when an error occurs, as indicated by the red cross in the right sub-figure, ***PyFi*-600K** enables us to trace the problem-solving chain from *DS* to *LR* and eventually to *PP*, thereby diagnosing how an error in the *DE* leads to the final mistake. Such a tracing-back approach is crucial for understanding the limitations of VLMs in financial decision-making, as it reveals at which capability level a failure occurs.

Identifying the Most Frequent Failure Capability Levels, as presented in the left part of Figure 4, reveals that the key to improving VLMs' performance in the financial domain lies in strengthening their capabilities at the CA (Calculation Analysis) level. Specifically, the pyramid-like structure and the *question chain* of ***PyFi-600K*** facilitates detailed error analysis, allowing us to identify specific levels where VLMs struggle. That is, most VLMs, even the advanced Claude-opus-4-1-20250805 and Qwen3-VL-Plus, fails to answer more than 40% of the question from CA, revealing that the calculation analysis is a critical bottleneck for them. More importantly, by comparing the error proportions between six capability levels, we can prioritize which capabilities to enhance first. For instance, since CA has the highest error proportion, focusing on improving calculation analysis skills in VLMs could lead to the most significant overall performance gains in financial decision-making tasks. In addition, this comparison also shows that, in the finance domain, analyzing numerical calculations from images is one of the most critical capabilities for accurate decision-making, even more important than logical reasoning, aligning with conclusions drawn by real-world finance experts.

Financial Theme: Economic Cycles & Market Theories - Level: Decision Support - Degree: 4



Question: Based on the image data on unemployment rates in Brazil (2012-2019), how does the high unemployment rate among the youth (15-24 years old) group impact a company's hiring strategy for new workers? **Image Background:** The image is Figure 1 titled 'Unemployment Rates and GDP, 2012-2019'. It is sourced from authors based on the National Household Sample Survey (Pesquisa Nacional por Amostra de Domicilios Continua, PNAD Continua) data from 2012 to 2019. The subject is unemployment rates and Gross Domestic Product (GDP) trends in Brazil over this period.

Qwen2.5-VL-3B-Instruct: The image shows that the unemployment rate among the youth (15-24 years old) group is consistently high throughout the period from 2012 to 2019. This high unemployment rate would likely decrease the availability of workers for hiring, making it unwise to hire from this group. Therefore, the company should not focus on hiring workers from this group. ✗

PyFi-QwenVL-3B: The line in the bottom graph that represents the unemployment rate for the youth group (15-24 years old) in Brazil is the yellow line.	*PP*
The blue line in the bottom graph of the image represents the unemployment rate for the total population in Brazil.	*PP*
The line in the bottom graph of the image that represents the unemployment rate for the youth group (15-24 years old) in Brazil is the yellow line.	*PP*
The approximate unemployment rate for the youth group (15-24 years old) in Brazil in 2019, as shown in the bottom graph, was 15%.	*DE*
The approximate unemployment rate for the youth group (15-24 years old) in Brazil in 2017, as shown in the bottom graph, was 10%.	*DE*
The approximate unemployment rate for the youth group (15-24 years old) in Brazil in 2014, as shown in the bottom graph, was 10%.	*DE*
The approximate percentage increase in the youth unemployment rate (15-24 years old) in Brazil from 2014 to 2019, based on the bottom graph, is 50%.	*CA*
The youth unemployment rate (15-24 years old) in Brazil increased from 2014 to 2019, based on the bottom graph.	*PR*
...	...
The youth unemployment rate (15-24 years old) in Brazil increased from 2014 to 2019, and this trend suggests that companies should prioritize hiring workers from this group to fill labor shortages.	*LR*
The final answer is: \\boxed{A}. ✓	*DS*

Figure 5: Qualitative analysis of level-wise COT by comparing the Qwen2.5-VL-3B-Instruct and *PyFi-QwenVL-3B* in the challenging financial decision-making.

Addressing Financial Decision-Making with Level-wise CoT, as depicted in Figure 5, is not only reliable but also highly interpretable. Specifically, we compare the responses of Qwen2.5-VL-3B-Instruct and *PyFi-QwenVL-3B* — the latter fine-tuned on *question chains*— when addressing the financial problem. The Qwen2.5-VL-3B-Instruct produces a short and disorganized reasoning process, reflecting its limited financial reasoning capability. In contrast, *PyFi-QwenVL-3B* follows a clear, structured, and logical reasoning flow, which is essential for expert-level performance in finance. For instance, our model first addresses basic *PP* questions, then proceeds to *DE* and *CA*, and finally tackles high-level *LR* and *DS*. This level-wise CoT not only improves answer reliability but also provides a transparent and traceable reasoning path that users can easily follow and verify, making it especially valuable in financial contexts where stakes are high. Moreover, experts can inspect this professional decision-making process and provide detailed and direct feedback to further refine the model.

Providing Process Supervision, in the form of a reward score for each sample in our ***PyFi-600K***, enables research into training reliable verifier models, known as process reward models (PRMs), to facilitate checking the correctness of each step generated by financial VLMs. Even though, due to space limitations, this paper does not perform such training, we use the reward scores to filter out low-quality samples during both evaluation and fine-tuning.

Mitigating Hallucinations. As shown in Figure 5 and Figure 4, our support of Level-wise CoT in training effectively mitigates model hallucination by enforcing a grounded reasoning process: high-level conclusions (e.g., DS) are derived strictly from verified low-level findings (e.g., PP and DE), rather than from ungrounded speculation. This structure ensures traceability and verifiability, as every reasoning step can be explicitly traced back to concrete evidence in preceding steps.

5 Conclusion

In this paper, we propose *PyFi*, a framework comprising the ***PyFi*-600K** dataset and the multi-agent adversarial mechanism *PyFi*-adv, to advance pyramid-like financial image understanding for VLMs. ***PyFi*-600K** features a pyramid-structured Q&A set, evaluating VLMs across 6 capability levels, 17 financial themes, and 11 image types. Its level-wise *question chain* enables interpretable financial decision-making by ranking capability importance and tracing lower-level errors. *PyFi*-adv allows scalable, annotation-free dataset growth. Experiments on VLMs and fine-tuned Qwen models show that *Calculation Analysis* is most critical for correct decisions, and fine-tuned small VLMs achieve a 19.52% accuracy gain without losing general task performance.

References

Anthropic. 2024. Claude 3 model family technical report.

Anthropic. 2026. Introducing claude opus 4.7. Accessed: 2026-04-16.

Rabiul Awal, Mahsa Massoud, Aarash Feizi, Zichao Li, Suyuchen Wang, Christopher Pal, Aishwarya Agrawal, David Vazquez, Siva Reddy, Juan A Rodriguez, and 1 others. 2025. Webmmu: A benchmark for multimodal multilingual website understanding and code generation. In *Proc 2025 Conference on Empirical Methods in Natural Language Processing*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Baidu-ERNIE-Team. 2025. Ernie 4.5 technical report.

Gagan Bhatia, Hasan Cavusoglu, Muhammad Abdul-Mageed, and 1 others. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13064–13087.

Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc.

Sijia Chen and Bin Song. 2025a. Rmultiplex200k: Toward reliable multimodal process supervision for visual language models on telecommunications. In *International Conference on Computer Vision (ICCV)*.

Sijia Chen and Bin Song. 2025b. Rmultiplex200k: Toward reliable multimodal process supervision for visual language models on telecommunications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1696.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024b. Autoprm: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *arXiv preprint arXiv:2402.11452*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proc 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Sanjiban Choudhury. 2025. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*.

SHUANGYAN DENG, Haizhou Peng, Jiachen Xu, Chunhou Liu, Ciprian Doru Giurcaneanu, and Jiamou Liu. 2025. Understanding financial reasoning in ai: A multimodal benchmark and error learning approach. In *ICLR Workshop on Reasoning and Planning for Large Language Models*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, pages 11733–11763. PMLR.

Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, and 1 others. 2025. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *ACM International Conference on Multimedia*.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Google. 2026. Gemini 3.5: frontier intelligence with action. Accessed: 2026-05-19.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Zhitao He, Zongwei Lyu, Dazhong Chen, Dadi Guo, and Yi R Fung. 2025. Matp-bench: Can mllm be a good automated theorem prover for multimodal problems? *arXiv preprint arXiv:2506.06034*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*, pages arXiv–2507.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Jimin Huang, Mengxi Xiao, Dong Li, Zihao Jiang, Yuzhe Yang, Yifei Zhang, Lingfei Qian, Yan Wang, Xueqing Peng, Yang Ren, and 1 others. 2024. Open-finllms: Open multimodal large language models for financial applications. *arXiv preprint arXiv:2408.11878*.

Seunghee Kim, Changhyeon Kim, and Taeuk Kim. 2024. Fcmr: Robust evaluation of financial cross-modal multi-hop reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Kai Lan, Jiayong Zhu, Jiangtong Li, Dawei Cheng, Guang Chen, and Changjun Jiang. 2025. Finlmm-r1: Enhancing financial reasoning in lmm through scalable data and reward design. *arXiv preprint arXiv:2506.13066*.

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

Jiangtong Li, Yiyun Zhu, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2025a. Cfbenchmark-mm: Chinese financial assistant benchmark for multimodal large language model. *arXiv preprint arXiv:2506.13055*.

Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, Zhiyong Huang, and Jing Ma. 2024. Mmcode: Benchmarking multimodal large language models for code generation with visually rich programming problems. In *Proc 2024 Conference on Empirical Methods in Natural Language Processing*.

Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2025b. Hedgeagents: A balanced-aware multi-agent financial trading system. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 296–305.

Xiaoyuan Li, Moxin Li, Wenjie Wang, Rui Men, Yichang Zhang, Fuli Feng, Dayiheng Liu, and Junyang Lin. 2025c. Mathopeval: A fine-grained evaluation benchmark for visual operations of mllms in mathematical reasoning. *arXiv preprint arXiv:2507.18140*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Longzhong Lin, Xuewu Lin, Kechun Xu, Haojian Lu, Lichao Huang, Rong Xiong, and Yue Wang. 2025. Revisit mixture models for multi-agent simulation: Experimental study within a unified framework. *arXiv preprint arXiv:2501.17015*.

Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. 2024a. Reasoning multi-agent behavioral topology for interactive autonomous driving. *Advances in Neural Information Processing Systems*, 37:92605–92637.

Wenhao Liu, Zhenyi Lu, Xinyu Hu, Jierui Zhang, Dailin Li, Jiacheng Cen, Huilin Cao, Haiteng Wang, Yuhan Li, Kun Xie, and 1 others. 2025a. Storm-born: A challenging mathematical derivations dataset curated via a human-in-the-loop multi-agent framework. In *arXiv preprint arXiv:2506.01531*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Junjie Lu, Zhaoling Chen, Chaofeng Qu, Jason Klein Liu, Chonghan Liu, Zefan Cai, Yunhui Xia, Li Zhao, Jiang Bian, and 1 others. 2025b. Adaptivestep: Automatically dividing reasoning step through model confidence. *arXiv preprint arXiv:2502.13943*.

Zhaowei Liu, Xin Guo, Fangqi Lou, Lingfeng Zeng, Jinyi Niu, Zixuan Wang, Jiajie Xu, Weige Cai, Ziwei Yang, Xueqian Zhao, and 1 others. 2025c. Fin-r1: A large language model for financial reasoning through reinforcement learning. *arXiv preprint arXiv:2503.16252*.

Zhaowei Liu, Xin Guo, Haotian Xia, Lingfeng Zeng, Fangqi Lou, Jinyi Niu, Mengping Li, Qi Qi, Jiahuan Li, Wei Zhang, and 1 others. 2025d. Visfineval: A scenario-driven chinese multimodal benchmark for holistic financial understanding. *arXiv preprint arXiv:2508.09641*.

Guilong Lu, Xuntao Guo, Rongjunchen Zhang, Wenqiao Zhu, and Ji Liu. 2025. Bizfinbench: A business-driven real-world financial benchmark for evaluating llms. *arXiv preprint arXiv:2505.19457*.

Jianqiao Lu, Zhiyang Dou, Hongru Wang, Zeyu Cao, Jianbo Dai, Yunlong Feng, and Zhijiang Guo. 2024a. Autopsv: Automated process-supervised verifier. *Advances in Neural Information Processing Systems*, 37:79935–79962.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, and 1 others. 2025. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. In *Annual Meeting of the Association for Computational Linguistics*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and 1 others. Improve mathematical reasoning in language models by automated process supervision, 2024. *URL 'https://arxiv.org/abs/2406.06592'*, 2.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *arXiv preprint arXiv:2310.10080*.

Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. 2025. Training on the benchmark is not all you need. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24948–24956.

OpenAI. 2025. Introducing gpt-4.1 model family. Accessed: 2025-10-02.

OpenAI. 2026. Introducing gpt-5.5. Accessed: 2026-04-23.

Hyejin Park, Jiyoon Lee, and Hayoung Oh. 2025. Fintab-llava: Finance domain-specific table understanding multimodal llm using fintmd. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 235–246. Springer.

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Jeff Zhao, Huan He, Yi Han, and 1 others. 2025. Multifinben: A multilingual, multimodal, and difficulty-aware benchmark for financial llm evaluation. *arXiv preprint arXiv:2506.14028*.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, and 1 others. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2025. Fin-fact: A benchmark dataset for multimodal financial fact-checking and explanation generation. In *Companion Proc ACM on Web Conference 2025*, pages 785–788.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proc 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

ShengbinYue ShengbinYue, Ting Huang, Zheng Jia, Siyuan Wang, Shujun Liu, Yun Song, Xuan-Jing Huang, and Zhongyu Wei. 2025. Multi-agent simulator drives language models for legal intensive interaction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6537–6570.

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, Haiyan Zhao, and Mengnan Du. 2025. Finchart-bench: Benchmarking financial chart comprehension in vision-language models. *arXiv preprint arXiv:2507.14823*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Xingwu Sun, Yanfeng Chen, Yiqing Huang, Ruobing Xie, Jiaqi Zhu, Kai Zhang, Shuaipeng Li, Zhen Yang, Jonny Han, Xiaobo Shu, and 1 others. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. 2024. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*.

Zichen Tang, Jiacheng Liu, Zhongjun Yang, Rongjin Li, Zihua Rong, Haoyang He, Zhuodi Hao, Xinyang Hu, Kun Ji, Ziyan Ma, and 1 others. 2025. Finmmr: Make financial numerical reasoning more multimodal, comprehensive, and challenging. In *International Conference on Computer Vision (ICCV)*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025b. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Hanbin Wang, Xiaoxuan Zhou, Zhipeng Xu, Keyuan Cheng, Yuxin Zuo, Kai Tian, Jingwei Song, Junting Lu, Wenhui Hu, and Xueyang Liu. 2025a. Codevision: Evaluating multimodal llms logic understanding and code generation capabilities. *arXiv preprint arXiv:2502.11829*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024b. Measuring multimodal mathematical reasoning with math-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peijie Wang, Zhong-Zhi Li, Fei Yin, Dekang Ran, and Cheng-Lin Liu. 2025b. Mv-math: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19541–19551.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024c. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proc Annual Meeting of the Association for Computational Linguistics*, pages 9426–9439.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2024d. Omnieval: An omnidirectional and automatic rag evaluation benchmark in financial domain. *arXiv preprint arXiv:2412.13018*.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025c. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025d. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023. Finvis-gpt: A multimodal large language model for financial chart analysis. *FinLLM Symposium at IJCAI 2023*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tengjin Weng, Jingyi Wang, Wenhao Jiang, and Zhong Ming. 2025. Visnumbench: Evaluating number sense of multimodal large language models. In *International Conference on Computer Vision (ICCV)*.

Chengyue Wu, Yixiao Ge, Qiushan Guo, Jiahao Wang, Zhixuan Liang, Zeyu Lu, Ying Shan, and Ping Luo. 2024a. Plot2code: A comprehensive benchmark for evaluating multi-modal large language models in code generation from scientific plots. *arXiv preprint arXiv:2405.07990*.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025a. Agentic reasoning: A streamlined framework for enhancing llm reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28489–28503.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, and 1 others. 2025b. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*.

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

Siyi Wu, Zhaoyang Guan, Leyi Zhao, Xinyuan Song, Xinyu Ying, Hanlin Zhang, Michele Pak, Yangfan He, Yi Xin, Jianhui Wang, and 1 others. 2025c. Mountainlion: A multi-modal llm-based agent system for interpretable and adaptive financial trading. *arXiv preprint arXiv:2507.20474*.

Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. 2025d. Visco: Benchmarking fine-grained critique and correction towards self-improvement in visual reasoning. In *Proc IEEE/CVF conference on computer vision and pattern recognition*, pages 9527–9537.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zixin Wang, Xinyi Xu, Wenxuan Wang, Zhiyao Xu, Yuhang Wang, and Michael R Lyu. 2024. Interaction2code: Benchmarking mllm-based interactive webpage code generation from interactive prototyping. *arXiv preprint arXiv:2411.03292*.

Jingyu Xiao, Ming Wang, Man Ho Lam, Yuxuan Wan, Junliang Liu, Yintong Huo, and Michael R Lyu. 2025. Designbench: A comprehensive benchmark for mllm-based front-end code generation. *arXiv preprint arXiv:2506.06251*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024.

Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484.

Vu Dinh Xuan, Hao Vo, David Murphy, and Hoang D Nguyen. 2025. Agentsgen: Multi-agent llm in the loop for semantic collaboration and generation of synthetic data. *arXiv preprint arXiv:2505.13466*.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2025. Famma: A benchmark for financial multilingual multimodal question answering. *arXiv preprint*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*.

Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, and Ruoming Pang. 2025b. Synthetic bootstrapped pretraining. *arXiv preprint arXiv:2509.15248*.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, and 1 others. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proc IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Yao Zhang, Yu Wu, Haowei Zhang, Weiguo Li, Haokun Chen, Jingpei Wu, Guohao Li, Zhen Han, and Volker Tresp. 2025a. Groundedprm: Tree-guided and fidelity-aware process reward modeling for step-level reasoning. *arXiv preprint arXiv:2510.14942*.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024b. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025b. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Suifeng Zhao, Zhuoran Jin, Sujian Li, and Jun Gao. 2025. Finragbench-v: A benchmark for multimodal rag with visual citation in the financial domain. *arXiv preprint arXiv:2505.17471*.

Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. 2025a. Fintmmbench: Benchmarking temporal-aware multi-modal rag in finance. In *ACM International Conference on Multimedia*.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. 2024. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix Contents

A Related Work Discussion . . . 15
A.1 VLM Reasoning . . . 15
A.2 Financial Benchmarks . . . 15
A.3 Automatic Data Augmentation . . . 16
A.4 Our Contributions . . . 16

B Data Details . . . 16
B.1 Data Statistics . . . 16
B.2 Instructions . . . 18
B.3 Data Processing . . . 19

C Limitations . . . 24

D Discussion on Data Leakage . . . 25

E Discussion on Pyramid-like Structure . . . 26

F Discussion on Question Chain . . . 26

G Model Performance with and without Fine-Tuning . . . 26

H Detailed Results . . . 36
H.1 Experiments Consideration and Ablation Studies . . . 36
H.2 Detailed Experiment Results on Original Evaluation Dataset with Image . . . 36

I Algorithms . . . 40

A Related Work Discussion

A.1 VLM Reasoning

Vision-language models (VLMs) have achieved impressive performance on complex tasks by leveraging step-by-step reasoning, decomposing intricate problems into simpler subtasks and solving them sequentially to reach a final solution (Guo et al., 2025; Team et al., 2025b; Bai et al., 2025; Hong et al., 2025; Team et al., 2025a; Wang et al., 2025d). Benchmarking reasoning ability is essential for guiding future advancements. In contrast to earlier methods that evaluated reasoning holistically via a single scalar score, recent work has shifted toward fine-grained assessment, enabling step-by-step validation of the chain-of-thought (CoT) process (Wu et al., 2025d; Chaoyou et al., 2023; Zhu et al., 2024; Chen et al., 2024a; Liu et al., 2024b; Yue et al., 2024; Chen and Song, 2025a). This paradigm has seen significant traction in domains like mathematical reasoning (Wang et al., 2024b; Lu et al., 2024b; Zhang et al., 2024a; Weng et al., 2025; Li et al., 2025c; He et al., 2025; Wang et al., 2025b; Qiao et al., 2024) and code generation (Xiao et al., 2025, 2024; Wang et al., 2025a; Awal et al., 2025; Li et al., 2024; Wu et al., 2024a), where outputs are inherently verifiable and intermediate steps can be automatically annotated.

The success in these domains stems from the inherent verifiability of their outputs, which enables automatic construction of process-annotated datasets for fine-tuning large language models (LLMs). Two types of verifiers are commonly employed: the Outcome Reward Model (ORM), which assigns a scalar reward to the entire solution, and the Process Reward Model (PRM) (Ma et al., 2023; Zhang et al., 2025b; Wang et al., 2025c; Choudhury, 2025; Lambert et al., 2024), which evaluates correctness at each step of the reasoning path. PRM's step-wise scoring enables precise error localization, facilitating more targeted feedback.

However, this progress has yet to be effectively transferred to vertical domains such as finance. Reasoning in finance typically involves multistep interpretation of charts, tables, and contextual narratives, requiring not only accurate understanding of financial visuals and deep domain knowledge, but also often yielding open-ended answers that are inherently difficult to verify precisely. This ambiguity impedes the generation of dense, step-wise process rewards essential for PRM training.

A.2 Financial Benchmarks

Existing financial benchmarks remain limited: they reduce complex reasoning to single scalar scores, obscure intermediate reasoning steps, and lack support for hierarchical or diagnostic evaluation (Kim et al., 2024; Zhu et al., 2025a; Shu et al., 2025; Li et al., 2025a; Liu et al., 2025d; Rangapur et al., 2025; Zhao et al., 2025; Gan et al., 2025; Luo et al., 2025; Xue et al., 2025; DENG et al., 2025). Additionally, current evaluation datasets consist only of isolated question-answer pairs, making it difficult to uncover hidden relationships within the data (Yang et al., 2025b). Compounding this challenge, these benchmarks rely on costly manual annotations that hinder scalability, and no large-scale synthetic dataset currently exists to train finance-specific preference reward models (PRMs) for visual reasoning tasks.

Specifically regarding scalability, FAMMA-LivePro (Xue et al., 2025), a manually annotated benchmark, contains 103 questions crafted by senior finance experts. These questions are of high quality and accuracy, and capture the complexity of real-world financial scenarios. However, such manual annotation demands highly qualified experts and is both time-consuming and labor-intensive, making it difficult to scale. While some benchmarks (Wang et al., 2024d; Gan et al., 2025; Luo et al., 2025; Lu et al., 2025; Li et al., 2025a; Zhu et al., 2025a; Zhao et al., 2025; Shu et al., 2025; Liu et al., 2024b) explore using large models for automatic data synthesis, they typically prompt these models to directly generate expert-level, complex financial questions. However, due to the difficulty large models face in understanding complex financial visuals, they often rely heavily on provided textual cues, leading to question homogenization and hallucinations. This results in question-answer pairs of inconsistent quality that still require substantial manual verification. On the other hand, most financial benchmarks only release their datasets without disclosing their data augmentation methods, making direct comparison of such methods difficult.

As a result, the finance domain still lacks benchmark datasets capable of supporting hierarchical capability assessment, evaluating step-by-step reasoning, enabling fine-grained reward modeling, and achieving fully scalable data augmentation.

A.3 Automatic Data Augmentation

Existing automated annotation methods fall into two categories. The first includes Math-Shepherd (Wang et al., 2024c), AutoPSV (Lu et al., 2024a), AutoPRM (Chen et al., 2024b), and AdaptiveStep (Liu et al., 2025b), which score or decompose steps of a fixed problem using verifiers or confidence heuristics. While efficient, they are passive and cannot generate new problems or structure them by difficulty. The second category, including OmegaPRM (Luo et al.), rStar-Math (Guan et al., 2025), GroundedPRM (Zhang et al., 2025a), and RMultiplex200K (Chen and Song, 2025b), leverages MCTS or structured planning to create higher-quality, verifiable paths. Specifically, OmegaPRM employs a divide-and-conquer MCTS algorithm to systematically locate the first incorrect step in a reasoning chain, enabling efficient and high-quality data generation. GroundedPRM also uses MCTS but integrates external validation tools to ensure the fidelity of each step, resulting in a more data-efficient PRM that achieves SOTA performance with far fewer annotations.

While existing methods such as Math-Shepherd, AutoPSV, AutoPRM, and AdaptiveStep effectively annotate steps of a given problem through passive, single-agent verification, they are inherently limited to flat, post-hoc labeling and cannot generate new problems or structure them by difficulty. Even MCTS-based approaches though capable of constructing high-fidelity reasoning paths—focus exclusively on verifying or refining a single solution trajectory. In contrast, our work introduces an active, adversarial two-agent framework embedded within the MCTS architecture: one agent (the questioner) generates financial reasoning problems, while the other (the answerer) solves them, with the questioner progressively increasing difficulty based on the answerer's performance.

This co-evolutionary process is specifically designed for the financial domain, where reasoning often requires layered knowledge, from basic accounting principles to complex risk modeling, and where errors must be interpretable and traceable. Our method explicitly constructs a hierarchical, interpretable curriculum of financial problems, in which each subsequent question builds upon prerequisite concepts and demands deeper domain understanding. Non-MCTS methods lack the structured search mechanism needed to jointly optimize problem generation, difficulty progression, and financial knowledge dependency. MCTS, by contrast, provides the necessary planning backbone to balance exploration of novel financial scenarios with exploitation of solvable, pedagogically meaningful chains. Thus, MCTS is essential not only for scalability but also for realizing a generative, knowledge-aware annotation process that produces explainable, difficulty-graded financial reasoning data, extending MCTS's role beyond verification to structured curriculum synthesis tailored to domain-specific reasoning.

A.4 Our Contributions

To address above shortcomings, we aim to: (1) comprehensively evaluate VLMs across a hierarchy of cognitive levels, from basic perception to expert financial decision-making, enabling granular diagnosis of strengths and weaknesses at each reasoning stage and revealing which foundational cognitions are critical for higher-level financial tasks, and uncover meaningful data relationships; (2) provide traceable, progressive logical chains that mirror real-world financial reasoning, enhancing interpretability, explainability, and justifiability while enabling VLMs to learn complex problem-solving through cumulative, step-by-step knowledge building; (3) introduce step-wise reward scores for each reasoning step, facilitating automated, scalable verification via Process Reward Models and overcoming the limitations of costly manual annotation, and (4) propose an automatic data augmentation method that decomposes the task of generating complex financial question-answer pairs into incremental process, starting from simple visual perception questions and progressively increasing difficulty and financial reasoning depth, all without requiring human verification. Finally, we present a comparison between our ***PyFi*-600K** dataset and existing evaluation datasets in Table 2.

B Data Details

B.1 Data Statistics

The complete ***PyFi*-600K** dataset comprises 603,859 samples, each comprising a financial image paired with a corresponding question-answer (Q&A) pair, along with essential background information (caption, theme, acronym expansions) of the image and a summary analysis of its context. Each sample is annotated with a capability level and complexity rating. Specifically, Level 1 (Perception) contains 73,006 samples (12.09% of the

Table 2: Comparison of ***PyFi*-600K** with Existing Evaluation Dataset. In Multimodal column, "T" means Text modality, "I" means Image modality. In the Volume column, K is used as the unit.

Dataset	Volume	Multimodal	Automatic Synthesized	Hierarchical Cognition	CoT	Step-wise Reward
General Domain						
MME (Chaoyou et al., 2023)	2.4	T & I			✓	
MMDocBench (Zhu et al., 2024)	4.3	T & I			✓	
MMStar (Chen et al., 2024a)	1.5	T & I				
MMBench (Liu et al., 2024b)	3.2	T & I		✓		
MMMU (Yue et al., 2024)	11.6	T & I				
Finance Domain						
FinQA (Chen et al., 2021)	8.3	T				
FLUE (Xie et al., 2023)	26.3	T				
FLARE (Xie et al., 2023)	19.2	T				
CF-Benchmark (Lei et al., 2023)	3.9	T				
BizFinBench (Lu et al., 2025)	7.0	T			✓	
FinMME (Luo et al., 2025)	11.1	T & I		✓		
MME-Finance (Gan et al., 2025)	1.2	T & I		✓		
FAMMA-Basic (Xue et al., 2025)	1.9	T & I			✓	
CFBenchmark-MM (Li et al., 2025a)	9.7	T & I				
FinMMR (Tang et al., 2025)	4.3	T & I			✓	
MultiFinBen-Vision (Peng et al., 2025)	21.8	T & I		✓		
***PyFi-600K* (Ours)**	600	T & I	✓	✓	✓	✓

Table 3: Question Length and Question Chain Length Statistics

Statistic	Question Length (Characters)	Chain Length
Mean	142.11	11.48
Median	137.00	12
Minimum	29	1
Maximum	853	26
Std Dev	50.54	2.55

Table 4: Distribution of Question Capability Levels and Complexity Degree

Classification		Type	Count	Percentage (%)
Capability Level	1	Perception	73,006	12.09
	2	Data Extraction	159,601	26.43
	3	Calculation Analysis	90,229	14.94
	4	Pattern Recognition	109,848	18.19
	5	Logical Reasoning	117,751	19.50
	6	Decision Support	53,424	8.85
Complexity Degree	1	—	91,143	15.09
	2	—	148,465	24.59
	3	—	165,072	27.34
	4	—	139,654	23.13
	5	—	59,527	9.86
Total			**603,859**	**100.00**

total); Level 2 (Data Extraction) includes 159,601 samples (26.43%); Level 3 (Calculation Analysis) comprises 90,229 samples (14.94%); Level 4 (Pattern Recognition) has 109,848 samples (18.19%); Level 5 (Logical Reasoning) accounts for 117,751 samples (19.50%); and Level 6 (Decision Support) includes 53,424 samples (8.85%). Notably, Level 2, Data Extraction, is the most prevalent, underscoring its foundational role in complex reasoning tasks. The dataset also includes over 60,000 sample chains, whose lengths vary from 1 to 26 samples, with an average chain length of 11.48. Question lengths range from 29 to 853 characters, with an average length of 142 characters, as shown in Table 3. The full distribution across capability levels and complexity degrees is shown in Table 4.

The distribution of question chain lengths by segment has shown in Table 5. The results indicate that the majority of sample chains (84.47%) fall within the 11-15 length range, suggesting that most tasks in the dataset require a moderate to substantial number of reasoning steps. Since chain length reflects the number of logically sequential question-answer pairs needed to arrive at a final answer, longer chains correspond to higher task complexity. The scarcity of very short (1-10, 15.11%) or very long (16-26, 0.42% combined) chains further implies that the dataset is intentionally curated to focus on multistep reasoning scenarios that are neither trivial nor excessively extended.

We also observed that, generally, the average question length increases with higher capability levels, indirectly reflecting a progressive rise in task difficulty, as shown in Figure 6. However, an unexpected decrease occurs at Level 4 (Pattern

Table 5: Distribution of question chain lengths by segment

Length Range	Count	Percentage (%)
1–10	9,456	15.11
11–15	52,863	84.47
16–20	241	0.39
21–26	21	0.03

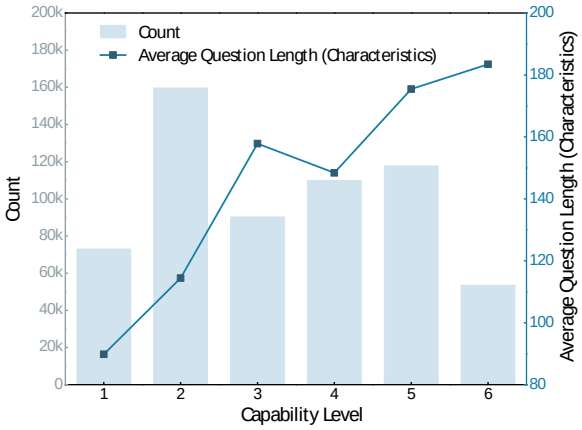


Figure 6: Capability Level with Question Length.

Table 6: Top most frequent words in the questions

Rank	Word	Frequency	Percentage (%)
1	price	112,502	1.31
2	growth	99,789	1.17
3	trend	83,465	0.97
4	gdp	63,960	0.75
5	debt	42,057	0.49
6	investment	38,931	0.45
7	income	34,923	0.41
8	share	33,044	0.39
9	revenue	29,665	0.35
10	change	29,019	0.34
11	risk	28,486	0.33
12	government	27,847	0.33
13	market	26,763	0.31

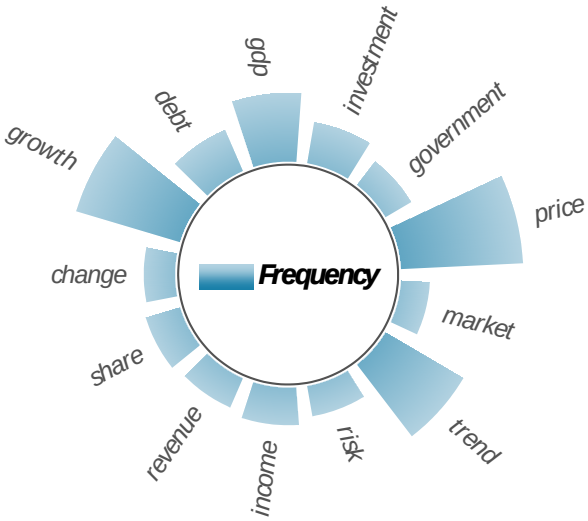


Figure 7: Top most frequent words in the questions.

Recognition). This may be because pattern recognition tasks often require identifying and interpreting intricate patterns rather than relying on simple textual matching, which could lead to more concise yet semantically dense questions that challenge model performance.

Finally, we analyzed word frequency in the questions, as shown in Table 6 and Figure 7. Excluding common function words such as prepositions and basic terms like "and" or "where," the most frequently occurring words include "price", "growth", "trend", "GDP", "debt", "investment", "income", "share", "revenue", "change", "risk", "government", and "market". This vocabulary highlights the dataset's rich thematic coverage across diverse domains of finance.

Before fine-tuning, we analyzed the answer distribution across all samples in the question chains and observed significant category imbalance. As shown in Table 7, answer option A accounts for 43.39% of all responses, while B makes up 34.20%. In contrast, options C and D represent only 15.81% and 6.08%, respectively, with options E-J collectively comprising less than 0.52%. This skew suggests a generation bias, likely introduced during synthetic question creation, where the model tends to place the correct answer in the first two choices. To mitigate this bias and ensure a fair evaluation, we manually rebalanced the answer distribution in a subset of 1,000 test samples so that options A, B, C, and D each appear with approximately equal frequency. During fine-tuning, we then compared model accuracy under the original imbalanced distribution versus the rebalanced version to quantify the impact of answer-position bias on model performance. This controlled comparison helps isolate whether apparent gains in accuracy stem from genuine reasoning ability or merely from exploiting skewed answer patterns.

B.2 Instructions

Data Privacy Data is collected from open-source and publicly available resources, including reports, textbooks, and online documents. When using MinerU (Wang et al., 2024a) for data extraction,

Table 7: Answer distribution in the Question Chain with non-standard responses grouped as "Wrong Format"

Answer Category	Percentage (%)
A	43.39
B	34.20
C	15.81
D	6.08
E	0.44
F	0.06
G	<0.01
H	<0.01
I	<0.01
J	<0.01
Wrong Format	<0.01
Total	**100.00**

we strictly adhere to the copyright and licensing terms of each source. Furthermore, we provide comprehensive citation information for every financial document in our collection, including details such as title, author, and publisher. Most importantly, the dataset is used exclusively for research purposes and involves no financial gain.

Data Format All textual data in ***PyFi*-600K** are formatted in JSON, while images are stored in `.jpg` format. We verify that each sample contains exactly one image, belongs to a specific level within the pyramid hierarchy, and is part of one or more sample chains.

Data Structure The complete dataset is stored in a tree structure represented in JSON format. The root node of the tree is a placeholder with no semantic meaning, while every other node represents a Q&A pair. Each branch extending from the root to a leaf node constitutes a complete sample chain. Nodes are assigned to different levels of a capability pyramid based on their hierarchical competence. Each node contains the following fields: the question, the answer, an associated image, essential visual context required to answer the question, and a summary/analysis of the contextual information.

B.3 Data Processing

To construct the final ***PyFi*-600K** dataset, we adopt a concise multi-stage pipeline: collect public financial documents; parse PDFs to extract images and associated text; evaluate compliance and classify images (normal/abnormal/extremely abnormal), enhancing only normal ones with contextual metadata; categorize images into content themes and chart types; maintain provenance; and sample by relevance and complexity with expert validation. An overview is shown in Figure 8.

We aggregated over 30,000 public financial PDFs in Chinese and English via web crawling and institutional APIs, then applied technical checks (corruption, password protection, readability) and domain relevance filters with curated inclusion/exclusion keywords to retain about 10,000 documents. We leverage `MinerU` (Wang et al., 2024a) to parse validated PDFs and extract images, text blocks, and layout metadata, preserving image–caption associations with page indices and bounding coordinates. Extracted assets were reorganized into a standardized hierarchical directory by document and page (separating images and contextual text) for efficient batch processing and traceability. We screened approximately 820,000 images using a VLM-guided dual 0–10 scoring for compliance (financial visualization criteria) and complexity (visual/data density and structural sophistication), generated per-image JSON summaries, and conducted a targeted manual audit. Images were labeled as normal/abnormal/extremely abnormal by combining caption-match windows around image tags with nearby-image proximity signals; only normal images were enhanced with contextual metadata. We follow the following criteria to select the target images for our study:

- Financial trend data chart only; no non-informative elements.
- Clear, legible Chinese or English text; unambiguous recognition.
- No obvious errors or omissions; follows financial data conventions.
- Clearly highlights dynamic financial processes.
- Tables/grids are secondary and must not dominate.
- Explicit relationships shown via visual connections.
- Addresses at least one finance-related issue.
- Universally discernible; self-contained composition.

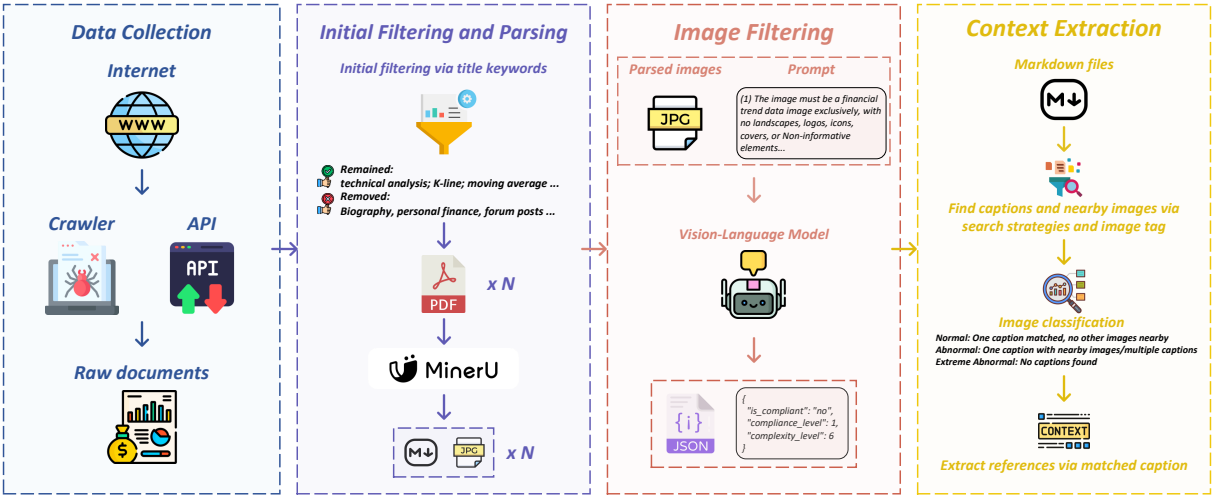


Figure 8: Overview of Data Processing Pipeline.

Context Extraction. The retained images are classified into three status categories (*normal*, *abnormal* and *extremely abnormal*) based on their contextual clarity within the document. A *normal* image is defined by a clear one-to-one match with a single caption located nearby and no other adjacent images, indicating reliable and unambiguous correspondence. Only images classified as *normal* are enriched with detailed contextual metadata from the surrounding text, providing each with explanatory insights.

An *abnormal* image arises when there is contextual ambiguity, either because the image has other images close to it or because multiple captions appear near it. An *extremely abnormal* image is identified when no caption is found within the search window, signaling a high risk of missing context. This classification ensures that contextual enhancement is applied only to images with a well-defined and correct caption relationship.

Topic Classification. We utilize the VLM to classify each image into one of 17 distinct financial content themes and one of 11 chart types, based on both the visual content of the image and its extracted contextual information.

Content Theme and Chart Type Classification

I. Content Theme

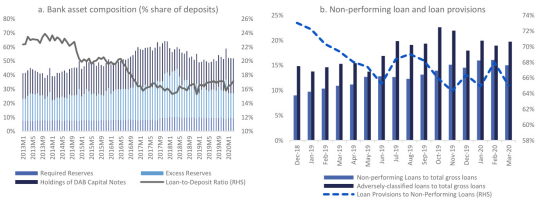
This dimension describes the core financial subject of the image, including *Macroeconomic Indicators*, *Financial Markets & Products*, *Commodities & Real Estate Markets*, *Bonds & Fixed Income*, *Monetary & Fiscal Policy*, *International Trade & Capital Flows*, *Corporate Finance & Valuation*, *Industry Analysis*, *Investment Theory & Portfolio Management*, *Risk Models & Management*, *Economic Cycles & Market Theories*, *Microeconomic Principles*, *Demographics & Socioeconomics*, *Financial Systems & Infrastructure*, *Geospatial Economic Data*, *Organization & Regulation*, and *Financial History & Documentation*, as shown in Figure 9 and Figure 10.

II. Chart Type

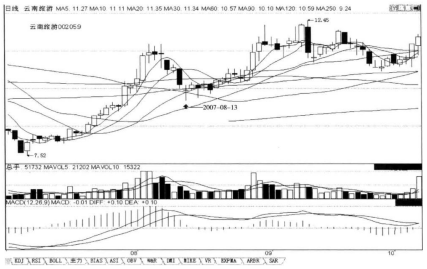
The chart types including *Line Chart*, *Bar Chart & Column Chart*, *Pie Chart & Donut Chart*, *Scatter Plot & Bubble Chart*, *Table*, *Diagram & Schematic*, *Radar Chart*, *Heatmap*, *Candlestick Chart*, *Photograph*, and *Infographic*, as shown in Figure 11.

Reference Info Extraction. We employ a LLM to extract bibliographic references and citation metadata from financial and academic documents. The model parses footnote, endnotes, and reference sections to identify cited works and extract key fields, including author(s), publication venue, year, and DOI (when available), while mapping each citation to its contextual use within the text.

Data Sampling. The image sampling strategy is structured around four key dimensions: complexity degree, compliance level, content theme, and chart type. The compliance level (ranging from 0 for non-compliant to 10 for fully compliant) reflects the degree to which the image satisfies the specified financial data visualization criteria. The complexity degree (from 0 for very simple to 10 for very complex) is assigned based on visual intricacy, data density, and structural sophistication. Each image

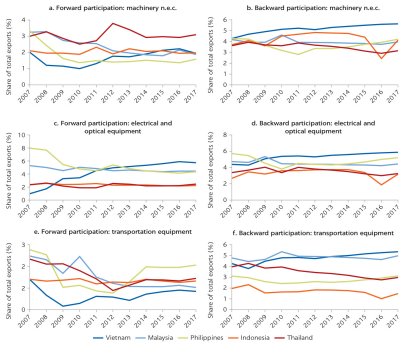


Macroeconomic Indicators

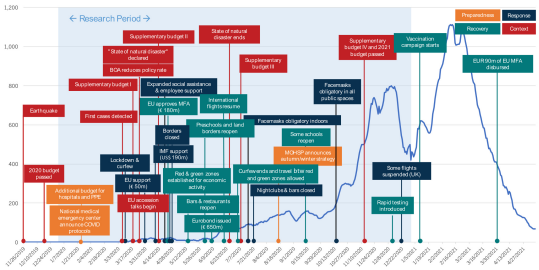


Financial Markets & Products

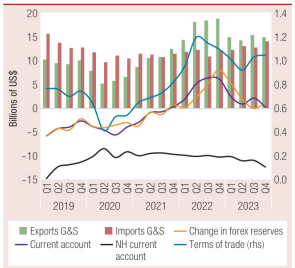
Commodities & Real Estate Markets



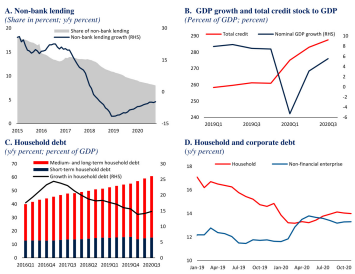
Industry Analysis



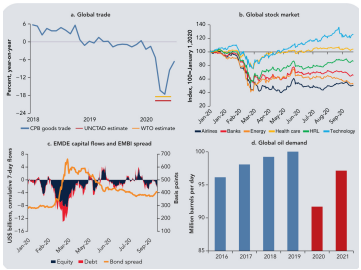
Monetary & Fiscal Policy



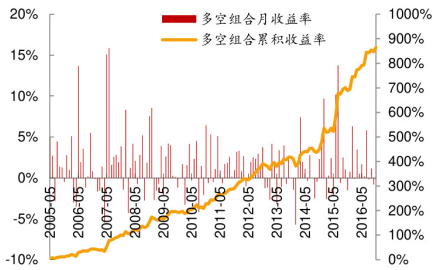
International Trade & Capital Flows



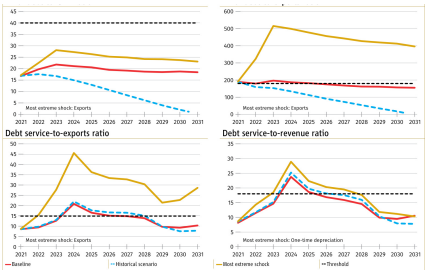
Corporate Finance & Valuation



Bonds & Fixed Income

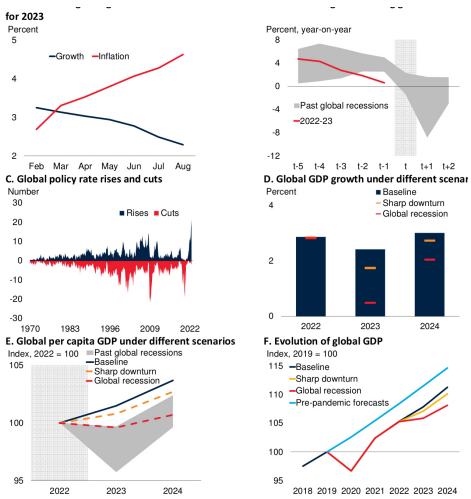


Investment Theory & Portfolio Management

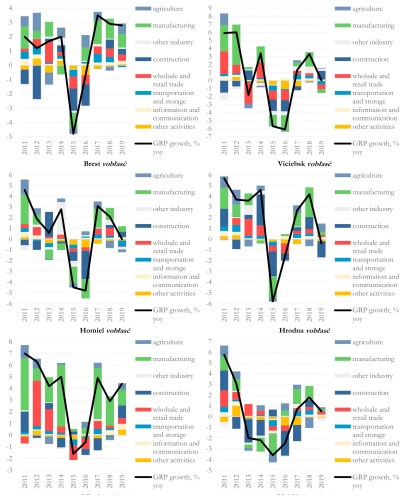


Risk Models & Management

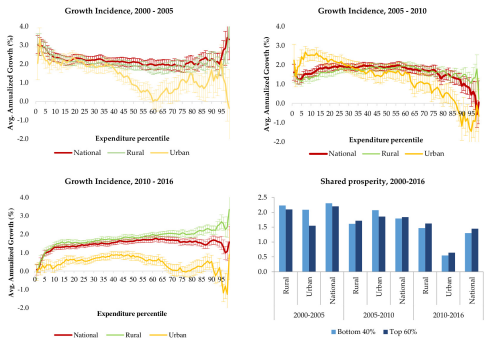
Figure 9: Overview of content themes.



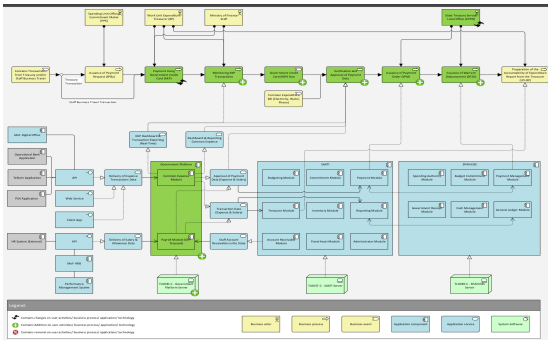
Economic Cycles & Market Theories



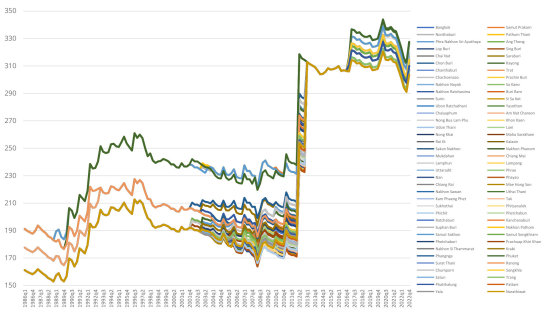
Geospatial Economic Data



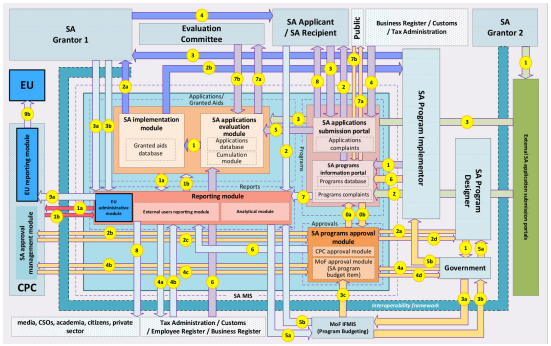
Demographics & Socioeconomics



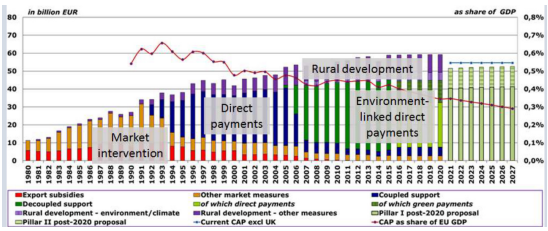
Financial Systems & Infrastructure



Microeconomic Principles

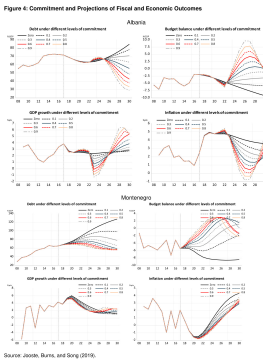


Organization & Regulation

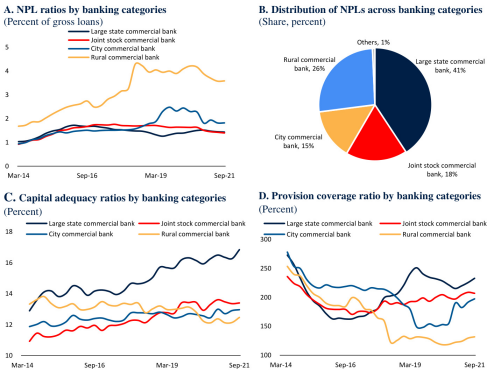


Financial History & Documentation

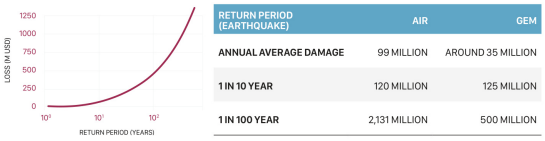
Figure 10: Overview of content themes (continued).



Line Chart

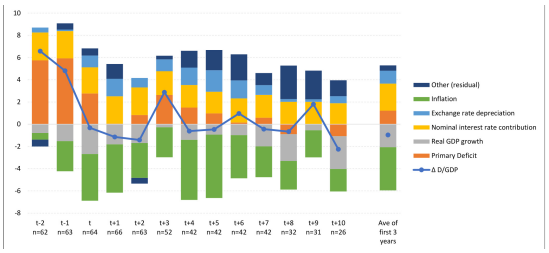


Pie Chart & Donut Chart

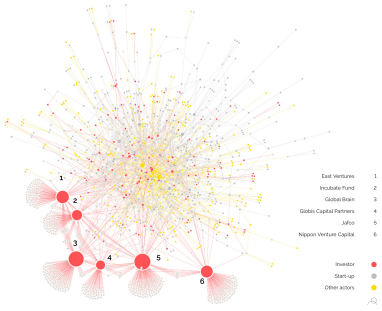


RETURN PERIOD (EARTHQUAKE)	AIR	GEM
ANNUAL AVERAGE DAMAGE	99 MILLION	AROUND 35 MILLION
1 IN 10 YEAR	120 MILLION	125 MILLION
1 IN 100 YEAR	2,131 MILLION	500 MILLION

Table

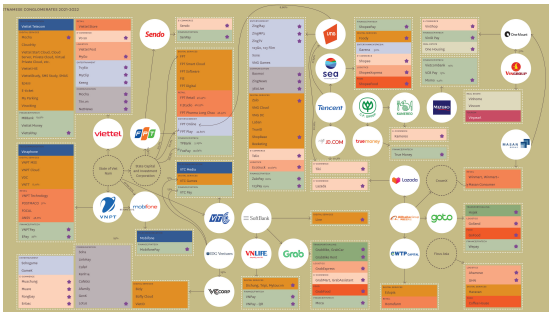


Bar Chart & Column Chart



Scatter Plot & Bubble Chart

Candlestick Chart



Diagram

Figure 11: Overview of chart types. In images used for generating question-answer pairs, Radar Charts, Heatmaps, and Photographs are rarely used.

is tagged with at least one of 17 content themes and at least one of 11 image types.

From an initial pool of **820,816** images, we applied a two-stage filtering process to construct a comprehensive financial evaluation dataset. First, we retained only images with compliance levels 9 and 10, indicating full or near-full adherence to the specified financial data visualization criteria. Non-financial images were excluded.

From the compliant subset, we selected the top

20,000 images with the highest complexity degree, reflecting visual intricacy and informational density. We further refined the set by including only image types relevant to financial data analysis: Line Chart, Bar Chart / Column Chart, Diagram / Schematic, Candlestick Chart / OHLC Chart, and Infographic. Content themes were balanced across categories such as macroeconomic indicators, financial markets, risk management, and economic cycles to ensure thematic diversity.

After automated filtering, a manual review was conducted to exclude any remaining unsuitable images, such as those with poor readability, non-compliant elements, or ambiguous content. The final curated dataset consists of **3,092** images. The resulting dataset supports rigorous computational analysis of financial visualizations with high compliance and varying complexity.

Initial Images Statistics Summary. A total of 820,816 images were collected and annotated from a substantial collection of financial documents. The initial dataset's distributions across content themes, image types, compliance levels, and complexity degrees are presented in Table 8.

Sampling Images Statistics Summary. A total of 3,092 images were selected from the initial dataset for analysis. The distribution across image types and themes is visualized in Figure 1 (c) and (d), respectively. Regarding the compliance level of the final dataset, 2,642 images (85.45%) were rated at level 9, while 450 images (14.55%) were rated at level 10. For complexity degree, the majority of images, 2,741 (88.65%), were rated at level 7, followed by 188 (6.08%) at level 8, 140 (4.53%) at level 6, 17 (0.55%) at level 9, and 6 (0.19%) at level 5.

C Limitations

We identify several key limitations in our current work:

Potential Data Contamination. Since ***PyFi*-600K** is derived from publicly available financial reports and books, there is a risk that these documents were included in the pre-training corpora of the evaluated VLMs. To address this, we implemented rigorous data-level filtering. We conducted "blind tests" to determine if questions could be answered correctly without visual input and "completion tests" to check if models could reconstruct content from text alone. Samples where models succeeded without visual context were flagged as potential leakage

Category	Classification	Count	Percentage
Content Theme	Macroeconomic Indicators	–	2.66%
	Financial Markets & Products	–	63.20%
	Commodities & Real Estate	–	0.99%
	Bonds & Fixed Income	–	0.73%
	Monetary & Fiscal Policy	–	0.88%
	International Trade	–	0.88%
	Corporate Finance & Valuation	–	1.39%
	Industry Analysis	–	2.98%
	Investment Theory & PM	–	13.42%
	Risk Models & Management	–	1.13%
	Economic Cycles & Theories	–	3.06%
	Microeconomic Principles	–	0.43%
	Demographics & Socioeconomics	–	3.57%
	Financial Systems	–	1.17%
	Organization & Regulation	–	1.35%
	Geospatial Economic Data	–	1.62%
	Financial History	–	0.53%
	Total	–	100.00%

(a) Content Theme Distribution

Category	Classification	Count	Percentage
Chart Type	Line Chart	–	11.82%
	Bar/Column Chart	–	5.52%
	Pie/Donut Chart	–	0.43%
	Scatter/Bubble Chart	–	0.45%
	Table	–	2.50%
	Diagram/Schematic	–	17.48%
	Radar Chart	–	0.07%
	Heatmap	–	0.05%
	Candlestick/OHLC	–	46.43%
	Photograph	–	3.99%
	Infographic	–	11.27%
	Total	–	100.00%

(b) Chart Type Distribution

Category	Classification	Count	Percentage
Compliance Level	1	194,325	23.67%
	2	3,703	0.45%
	3	995	0.12%
	4	9	0.00%
	5	117	0.01%
	6	614	0.07%
	7	34,455	4.20%
	8	561,368	68.39%
	9	23,206	2.83%
	10	2,024	0.25%
	Total	820,816	100.00%

(c) Compliance Level Distribution

Category	Classification	Count	Percentage
Complexity Degree	0	35	0.00%
	1	52,841	6.44%
	2	49,852	6.07%
	3	49,937	6.08%
	4	49,743	6.06%
	5	82,867	10.10%
	6	408,909	49.82%
	7	126,258	15.38%
	8	296	0.04%
	9	78	0.01%
	Total	820,816	100.00%

(d) Complexity Degree Distribution

Table 8: Initial Image Distributions

and excluded from our core evaluation set to ensure that performance gains reflect genuine multimodal reasoning. Although it may always be beneficial to include documents published after the release of large language models, doing so can significantly reduce the scalability and size of the dataset, introduce manual effort, and prevent the data from being updated continuously.

Computational Constraints on Scale. Although ***PyFi*-600K** contains 600k samples, computational limits restricted us from scaling further. However, the pipeline is designed to be extensible and can leverage the vast array of financial datasets available (e.g., on Hugging Face) to construct billion-scale datasets in the future.

Reward Modeling Noise. The reward scores used during the adversarial interactions and MCTS may contain noise. Accurately evaluating open-ended financial reasoning is inherently challenging, and minor inaccuracies in the reward model can affect the optimization process.

Adversarial Framework Complexity. The *PyFi-adv* mechanism, while effective for data synthesis, introduces complexity that may limit its deployment in resource-constrained environments. Furthermore, the stochastic nature of VLMs can occasionally lead to inconsistent chain lengths or hallucinations, which requires careful post-hoc filtering.

Incomplete Process Supervision Analysis. Due to time and resource limitations, we have not yet fully explored the upper bounds of process supervision provided by our framework, although our initial experiments demonstrate its significant value.

D Discussion on Data Leakage

In constructing ***PyFi*-600K**, we acknowledge the potential risk of **data leakage**, particularly due to the use of publicly available financial documents that may have been included in the pretraining corpora of some VLMs. To ensure a fair and reliable evaluation, we conduct a **systematic and stringent** analysis to identify and mitigate such leakage.

Existing research employs various methods to detect data leakage. Dong et al. (2024) identify leakage by measuring the peakedness of the model's output distribution, where high similarity among sampled outputs indicates memorization. Wu et al. (2025b) assess leakage by the model's ability to reconstruct or answer problems based only on partial prompt prefixes. Ni et al. (2025) detect leakage in multiple-choice benchmarks by identifying if the log-probability of the original option order is a significant outlier among permutations. However, these approaches are often indirect, model-specific, or limited to particular benchmark formats, making them less applicable to financial image understanding tasks where the interplay between image and text is essential.

We define a sample as exhibiting data leakage if **more than half of the evaluated models** can answer the question correctly **without access to the image**. This criterion is based on the assumption that if models can consistently answer a question without visual input, the answer is likely encoded in the textual content or has been memorized during pretraining, rather than being inferred from the image. **Our method is more direct, task-agnostic, and interpretable**: by explicitly ablating the visual modality and measuring performance across diverse models, we capture leakage that genuinely compromises multimodal evaluation fairness—making it the most suitable strategy for benchmarks like *PyFi*-600K that hinge on visual grounding.

To quantify this, we conduct a comprehensive study where we evaluate all 15 VLMs on the original 1,000-sample test set under two conditions: with and without the image input. Samples that meet the leakage criterion are subsequently excluded from the evaluation set. As a result, the test set is reduced from 1,000 to **301 samples**, forming the **Evaluation Dataset after Exclusion**. **All results reported in the main paper are derived exclusively from this rigorously filtered set.**

This filtered dataset is used for all final evaluations and fine-tuning experiments reported in the main paper. The significant reduction in sample size underscores the prevalence of potential leakage in financial Q&A datasets and highlights the importance of this filtering step.

This approach is both **reasonable and effective** in mitigating the influence of data leakage. By removing samples that can be answered without visual reasoning, we ensure that the remaining questions require genuine visual understanding and financial reasoning, aligning with the goals of ***PyFi*-600K**. Furthermore, this method is model-agnostic and does not rely on external knowledge of pretraining data, making it broadly applicable.

In summary, our data leakage mitigation strategy enhances the validity of the evaluation results and supports the development of more robust and generalizable financial VLMs.

E Discussion on Pyramid-like Structure

Financial image analysis requires more than object detection or simple classification; true understanding is a multi-stage cognitive process from visual perception to strategic decision-making. To address this, we propose ***PyFi*-600K**, a pyramid-like hierarchy structured across six levels: **Perception**, **Data Extraction**, **Calculation Analysis**, **Pattern Recognition**, **Logical Reasoning**, and **Decision Support**.

Level 1: Perception. This is the foundational base of the pyramid. At this stage, the model performs basic visual recognition. It identifies fundamental elements within the image: "What objects are present?" or "What is the color of that line?" This involves detecting axes, labels, data points, lines, bars, pie segments, and textual annotations. The questions and answers are a structured inventory of visual components.

Level 2: Data Extraction. Building on the perceived components, this level translates visual information into machine-readable quantitative or categorical data. For example, it converts the height of a bar in a chart into a numerical value based on the Y-axis scale, or it extracts the exact numerical percentage from a pie chart label. The output is a clean dataset ready for computation.

Level 3: Calculation Analysis. Here, the extracted data undergoes initial processing. This involves performing basic arithmetic, such as calculating percentage changes, growth rates or variances, and descriptive statistics. This level answers "What are the basic quantitative relationships?"

Level 4: Pattern Recognition. This level moves from discrete calculations to identifying broader trends and shapes within the data. It answers "What is happening over time or across categories?" Examples include recognizing an upward or downward trend in a stock price chart, identifying a period of high volatility, or spotting seasonal patterns in sales data.

Level 5: Logical Reasoning. This is a critical step towards higher-order cognition. The model integrates the identified patterns with external or common-sense financial knowledge to answer "Why is this happening, or what does this imply?" For instance, reasoning that a sharp drop in a company's stock price coinciding with a negative earnings report indicates a causal relationship. It involves drawing inferences and establishing cause-and-effect links.

Level 6: Decision Support. The apex of the pyramid leverages all previous levels to provide actionable insights. It answers "What should be done?" Based on the reasoned analysis and the options or answers might include recommendations.

***PyFi*-600K** provides a robust and principled framework for building a multi-modal financial image understanding dataset. By decomposing the task into six hierarchically organized cognitive levels for automated, reward-driven generation of logical chains, we create a resource that promotes deep, structured, and explainable reasoning.

F Discussion on Question Chain

A complete question chain can refer to the Figure 12, Figure 13, and Figure 14.

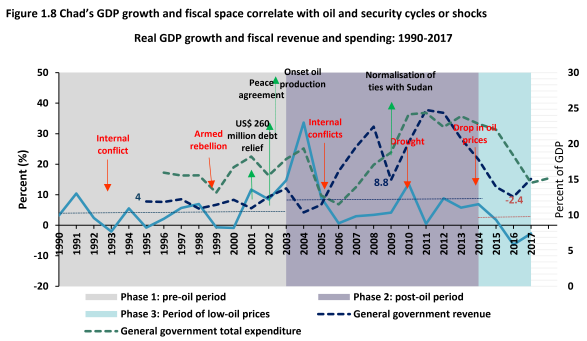


Figure 12: Illustration of the complete question chain presented in the Figure 2 of the main paper.

G Model Performance with and without Fine-Tuning

The examples of Qwen2.5-VL-3B-Instruct performance before and after fine-tuning can refer to the Figures 15, 16, and 17, and the examples of Qwen2.5-VL-7B-Instruct can refer to the Figures 18, 19, 20 and 21.

Complete Example of Question Chain

1. *Question:* Which shaded region in the chart represents the 'Period of low-oil prices'?
Options:
A. Gray (Phase 1: pre-oil period)
B. Purple (Phase 2: post-oil period)
C. Light blue (Phase 3: Period of low-oil prices)
D. Green (not a shaded region)
Answer: **C**

2. *Question:* What is the color of the shaded region labeled 'Phase 2: post-oil period' in the chart?
Options:
A. Light gray
B. Dark gray
C. Light blue
D. Cyan
Answer: **B**

3. *Question:* During Phase 3: Period of low-oil prices, what is the approximate value of general government revenue (blue dashed line) in Chad, as shown in the chart?
Options:
A. 5% of GDP
B. -2.4% of GDP
C. 10% of GDP
D. 20% of GDP
Answer: **B**

4. *Question:* Which event is associated with the decline in general government revenue (blue dashed line) during Phase 3 (Period of low-oil prices) in Chad, as shown in the chart?
Options:
A. Internal conflict
B. Drought
C. Drop in oil prices
D. Normalisation of ties with Sudan
Answer: **C**

5. *Question:* In 2016, during the Period of low-oil prices (Phase 3), what is the approximate value of (general government total expenditure - general government revenue) in Chad?
Options:
A. 17.4%
B. -17.4%
C. 5%
D. -5%
Answer: **A**

6. *Question:* What pattern is observed between general government revenue and total expenditure during the 'Onset oil production' period (marked in the chart)?
Options:
A. Both revenue and expenditure increase significantly
B. Revenue increases while expenditure decreases
C. Revenue decreases while expenditure increases
D. Both revenue and expenditure remain stable
Answer: **A**

Figure 13: Complete example of question chain.

Complete Example of Question Chain (continued)

7. *Question:* What pattern does the general government total expenditure (green dashed line) follow during periods of internal conflict (marked by red arrows) in Chad from 1990 to 2017?
Options:
A. It decreases consistently
B. It increases significantly
C. It remains stable
D. It fluctuates randomly
Answer: **B**

8. *Question:* How does the volatility of general government revenue in Chad during the Period of low-oil prices (Phase 3) contribute to economic vulnerability?
Options:
A. It increases military spending, leading to internal conflict.
B. It reduces the government's ability to withstand shocks and invest in stability.
C. It causes oil prices to drop further, worsening revenue.
D. It eliminates the need for non-oil revenue sources.
Answer: **B**

9. *Question:* Based on the image illustrating Chad's economic volatility driven by oil price cycles and security shocks, how would establishing a sovereign wealth fund help reduce revenue vulnerability during low-oil price periods?
Options:

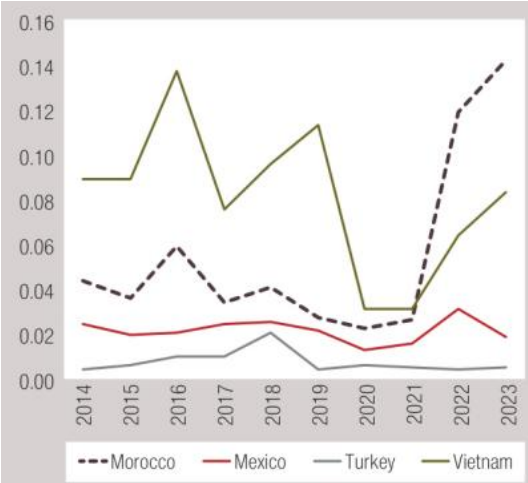
A. By increasing military spending to preempt conflict and stabilize revenue
B. By investing oil revenues during high-price periods to fund expenditures during low-price periods
C. By reducing reliance on non-oil revenue sources, such as IMF-ECF support
D. By increasing general government revenue through borrowing from the World Bank during booms
Answer: **B**

10. *Question:* Which policy would most effectively address both Chad's oil revenue volatility (from external shocks) and vulnerability to internal conflict?
Options:
A. Increase military spending to preempt conflict
B. Establish a sovereign wealth fund to stabilize revenues and investments
C. Reduce all non-oil revenue sources to avoid over-reliance on oil
D. Borrow heavily from the World Bank during low-oil price periods
Answer: **B**

11. *Final Question:* Based on the image illustrating Chad's fiscal volatility driven by oil cycles and security shocks, which policy would most effectively reduce economic vulnerability to future external shocks and internal conflict?
Options:
A. Increase military spending to preempt conflict
B. Establish a sovereign wealth fund to stabilize revenues and investments
C. Reduce all non-oil revenue sources to avoid over-reliance on oil
D. Borrow heavily from the World Bank during low-oil price periods
Answer: **B**

Figure 14: Complete example of question chain (continued).

International Trade & Capital Flows - Level: Decision Support - Degree: 4



Question:
From 2022 to 2023, which country had the highest FDI announcements as a share of GDP growth rate, based on the chart?

Options:
A: Morocco
B: Mexico
C: Turkey
D: Vietnam

Image Background:
The image is Figure 11 titled 'FDI Morocco outperformed peers (FDI announcements as a share of GDP)'. It is sourced from FDI Markets, a service from the Financial Times Ltd., and Office des Changes, with calculations and analyses by IFC. The image path is [images/002131/000013.jpg].

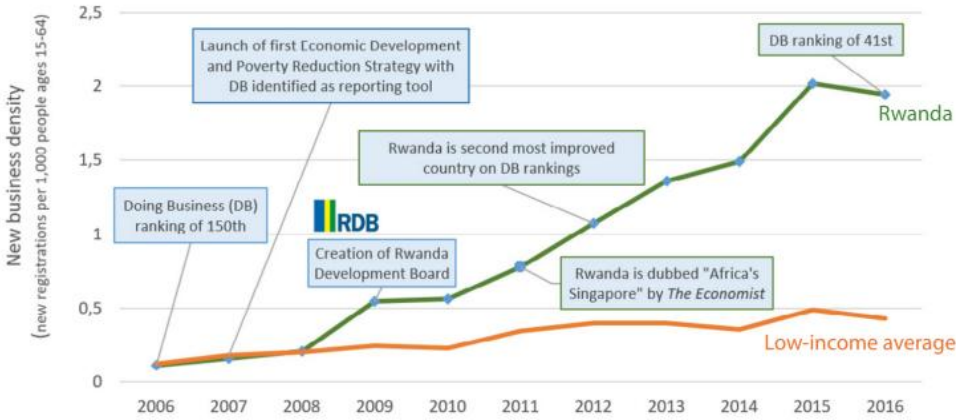
Qwen2.5-VL-3B-Instruct:
To determine which country had the highest FDI announcements as a share of GDP growth rate from 2022 to 2023, we need to look at the trend lines for each country in the chart. The chart shows four countries: Morocco, Mexico, Turkey, and Vietnam. We are interested in the period from 2022 to 2023.
- Morocco's trend line shows a significant increase from 2022 to 2023.
- Mexico's trend line shows a slight increase from 2022 to 2023.
- Turkey's trend line shows a slight decrease from 2022 to 2023.
- Vietnam's trend line shows a slight increase from 2022 to 2023.

Given this information, Morocco's trend line shows the most significant increase from 2022 to 2023, indicating the highest growth rate among the four countries.

PyFi-QwenVL-3B:
The country represented by the red line in the chart is Mexico. *PP*
The country represented by the green line in the chart is Vietnam. *PP*
The country represented by the dashed brown line in the chart is Morocco. *PP*
The approximate FDI announcements as a share of GDP for Morocco in 2022, based on the chart, is 0.12. *DE*
The approximate FDI announcements as a share of GDP for Vietnam in 2022, based on the chart, is 0.08. *DE*
The approximate FDI announcements as a share of GDP for Mexico in 2022, based on the chart, is 0.03. *DE*
The approximate percentage growth rate of FDI announcements as a share of GDP for Morocco from 2021 to 2022, based on the chart, is 200%. *CA*
The approximate percentage growth rate of FDI announcements as a share of GDP for Vietnam from 2021 to 2022, based on the chart, is 150%. *CA*
The country whose FDI announcements as a share of GDP line shows the most significant upward trend from 2021 to 2022 is Morocco. *CA*
The country with the highest FDI announcements as a share of GDP growth rate from 2021 to 2022, based on the chart, is Morocco. *PR*
The country with the highest FDI announcements as a share of GDP growth rate from 2022 to 2023, based on the chart, is Morocco. *PR*
The approximate percentage growth rate of FDI announcements as a share of GDP for Turkey from 2021 to 2022, based on the chart, is 0%. *LR*
The approximate percentage growth rate of FDI announcements as a share of GDP for Mexico from 2021 to 2022, based on the chart, is 50%.The final answer is: \\boxed{A} *DS*

Figure 15: Model performance before and after fine-tuning (Qwen2.5-VL-3B-Instruct)

Organization & Regulation - Level: Decision Support - Degree: 5



Question:

Which factor most directly led to Rwanda's new business density exceeding the low-income average after 2009?

Options:

A: Implementation of Rwanda Development Board (RDB) and using Doing Business (DB) as a reporting tool.

B: A rapid improvement in Doing Business (DB) ranking from 150th to 41st.

C: Natural economic growth unrelated to policy reforms.

D: Increased foreign direct investment (FDI) without regulatory changes

Image Background:

The image is titled 'Figure 2. A timeline of Rwanda's business reform successes'. It visually represents Rwanda's business regulatory reforms from 2009 (Doing Business 2010) to 2018 (Doing Business 2019), implemented under the Rwanda Development Board (RDB), which was established in 2009. The RDB oversees business regulations, foreign direct investment (FDI), and broader development planning, inspired by the Singaporean model. The source of the data is the Doing Business data from the World Bank Entrepreneurship database. Acronyms and abbreviations in the image: RDB = Rwanda Development Board;

Qwen2.5-VL-3B-Instruct:

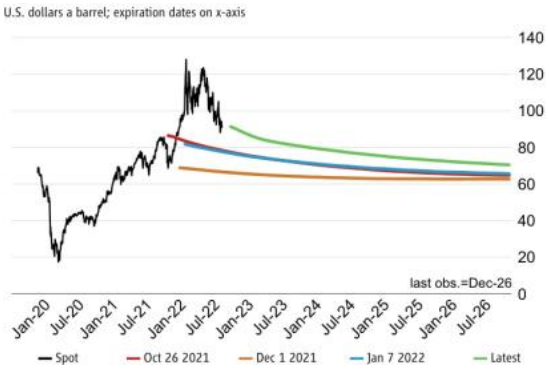
The image shows a timeline of Rwanda's business regulatory reforms from 2009 to 2018, with the Rwanda Development Board (RDB) being established in 2009. The timeline indicates that Rwanda's Doing Business (DB) ranking improved significantly from 150th to 41st, which is a major factor in the country's business density exceeding the low-income average. The RDB's role in overseeing business regulations and foreign direct investment (FDI) is also highlighted, suggesting that policy reforms were a key driver of the business density increase.

PyFi-QwenVL-3B:

The color line in the chart that represents Rwanda's new business density is green. *PP*
The color line in the chart that represents the low-income average new business density is orange. *PP*
The year when Rwanda's Doing Business (DB) ranking first exceeded 150th is 2009. *DE*
The year when Rwanda's Doing Business (DB) ranking first exceeded 150th, as indicated by the chart, is 2009. *DE*
The approximate percentage increase in Rwanda's new business density from 2009 to 2015, based on the chart, is 150%. *CA*
The approximate percentage increase in Rwanda's new business density from 2015 to 2016, based on the chart, is 50%. *CA*
The factor most directly associated with Rwanda's new business density exceeding the low-income average after 2009, based on the chart and contextual information, is the implementation of Rwanda Development Board (RDB) and using Doing Business (DB) as a reporting tool. *PR*
The factor most directly associated with Rwanda's Doing Business (DB) ranking improvement from 2009 to 2015, based on the chart and contextual information, is the creation of Rwanda Development Board (RDB). *LR*
... ...
The factor most directly associated with Rwanda's new business density exceeding the low-income average after 2009, based on the chart and contextual information, is the creation of Rwanda Development Board (RDB) and using Doing Business (DB) as a reporting tool. *LR*
The final answer is: \\boxed{A}. *DS*

Figure 16: Model performance before and after fine-tuning (Qwen2.5-VL-3B-Instruct)

Commodities & Real Estate Markets - Level: Decision Support - Degree: 4



Question:

If a trader has a long position in a near - expiry (3 - month) oil futures contract and the longer - dated (6 - month) contract is cheaper, what is the impact on their potential profit if they roll over the position?

Options:

A: Profit decreases because the new contract is cheaper.

B: Profit decreases because the new contract is more expensive.

C: Profit increases because the new contract is cheaper.

D: Profit increases because the new contract is more expensive.

Image Background:

The image is titled 'Figure 2.1. Oil and natural gas prices' and contains two panels: Panel A shows Brent Oil spot prices, and Panel B shows European-delivery natural gas prices. The image is referenced in the context of discussing natural gas price trends in Europe amid geopolitical tensions related to the war in Ukraine. Abbreviations: MENA (Middle East and North Africa), GCC (Gulf Cooperation Council), EU (European Union), IMF (International Monetary Fund), WEO (World Economic Outlook), UN Comtrade (United Nations Commodity Trade Statistics Database), G20 (Group of Twenty), AES (Advanced Economies in the G20: Australia, Canada, Eurozone, France, Germany, Italy, Japan, Saudi Arabia, the United Kingdom, the United States), EMs (Emerging Markets in the G20: Brazil, China, India, Mexico, South Africa, South Korea, and Turkey).

Qwen2.5-VL-3B-Instruct:

The trader has a long position in a near-expiry (3-month) oil futures contract. If the longer-dated (6-month) contract is cheaper, rolling over the position would mean the trader would switch to the cheaper contract. This would result in a decrease in the trader's potential profit because the new contract is cheaper, meaning the trader would have to pay less for the same quantity of oil. Therefore, the trader's profit decreases.

PyFi-QwenVL-3B:

The color line in Panel A that represents the latest Brent Oil spot price is Green. *PP*

The color line in Panel A that represents the 6-month Brent Oil futures contract is Blue. *PP*

The color line in Panel A that represents the 3-month Brent Oil futures contract is Red. *PP*

The approximate price of the 3-month Brent Oil futures contract (red line) in Panel A at its latest expiration date (last observed date) is 60 U.S. dollars per barrel. *DE*

The approximate price of the 6-month Brent Oil futures contract (blue line) at its latest expiration date (last observed date) in Panel A is 65 U.S. dollars per barrel. *DE*

The approximate price of the 3-month Brent Oil futures contract (red line) at its latest expiration date (last observed date) in Panel A is 60 U.S. dollars per barrel. *DE*

The approximate percentage increase in the price of the 6-month Brent Oil futures contract from its latest expiration date to its 3-month expiration date is 8.3%. *PR*

The approximate percentage decrease in the price of the 3-month Brent Oil futures contract from its latest expiration date to its 6-month expiration date is 16.7%. *PR*

The price trend of the 3-month Brent Oil futures contract (red line) from its latest expiration date to its 6-month expiration date is decreasing. *PR*

The price trend of the 6-month Brent Oil futures contract (blue line) from its latest expiration date to its 3-month expiration date is increasing. *PR*

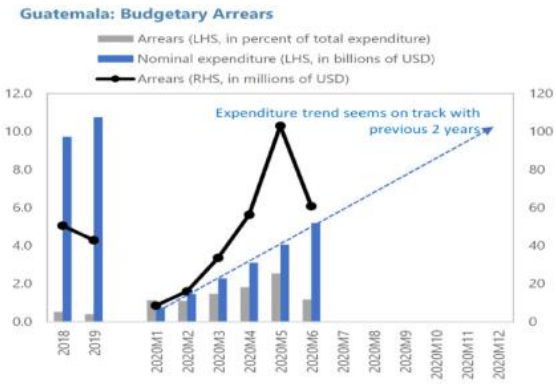
If a trader has a long position in a 3-month Brent Oil futures contract and the 6-month contract is cheaper, the trader's profit from rolling over the position decreases because the new contract is cheaper. *LR*

If a trader has a long position in a 3-month Brent Oil futures contract and the 6-month contract is cheaper, the trader's profit from rolling over the position decreases because the new contract is cheaper. *LR*

If a trader has a long position in a 3-month Brent Oil futures contract and the 6-month contract is cheaper, the trader's profit from rolling over the position decreases because the new contract is cheaper. The final answer is: \\boxed{A} *DS*

Figure 17: Model performance before and after fine-tuning (Qwen2.5-VL-3B-Instruct)

Monetary & Fiscal Policy - Level: Decision Support - Degree: 4



Question:

Which combination of trends in the chart shows that Guatemala avoided fiscal collapse during the COVID-19 crisis?

Options:

A: Nominal expenditure (blue bars) continued rising while arrears (black line) peaked then declined

B: Nominal expenditure (blue bars) declined while arrears (gray bars) remained low

C: Nominal expenditure (blue bars) fell sharply while arrears (black line) decreased

D: Nominal expenditure (blue bars) fluctuated wildly while arrears (gray bars) stayed constant

Image Background:

The image is labeled 'FIGURE 23 - Guatemala: Budgetary Arrears' and appears in the context of a discussion on government arrears in Guatemala during the COVID-19 crisis. The subject is Guatemala's budgetary arrears, with reference to the World Bank (WB) and International Monetary Fund (IMF).

Qwen2.5-VL-7B-Instruct:

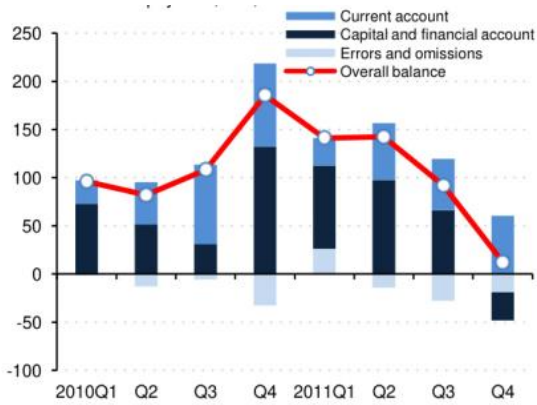
To determine which combination of trends in the chart shows that Guatemala avoided fiscal collapse during the COVID-19 crisis, we need to analyze the behavior of nominal expenditure and arrears over time.
1. **Nominal Expenditure (blue bars)**: The blue bars represent the nominal expenditure in billions of USD. We need to observe if there was a significant decline or increase in expenditure.
2. **Arrears (black line)**: The black line represents the arrears in millions of USD. We need to observe if the arrears peaked and then declined, indicating that the government was able to manage its financial obligations.
From the chart:
- The blue bars (nominal expenditure) show a general increase over time, which is consistent with the trend of previous years.\n- The black line (arrears) shows a peak in 2020M5 and then declines in subsequent months, indicating that the government was able to manage its financial obligations.
Given these observations, the combination of trends that shows Guatemala avoided fiscal collapse is:
- Nominal expenditure (blue bars) continued rising while arrears (black line) peaked then declined.
Therefore, the correct answer is A.

PyFi-QwenVL-7B:

The unit of measurement for the black line in the chart labeled 'Guatemala: Budgetary Arrears' is millions of USD. *PP*
The unit of measurement for the gray bars in the chart labeled 'Guatemala: Budgetary Arrears' is percent of total expenditure. *PP*
The unit of measurement for the blue bars in the chart labeled 'Guatemala: Budgetary Arrears' is billions of USD. *PP*
The approximate value of nominal expenditure (blue bars) in 2020M5 (May 2020) was 4 billion USD. *DE*
The approximate value of nominal expenditure (blue bars) in 2020M1 (January 2020) was 1 billion USD. *DE*
The approximate value of nominal expenditure (blue bars) in 2020M9 (September 2020) was 5 billion USD. *DE*
The approximate percentage increase in nominal expenditure (blue bars) from 2020M1 (January 2020) to 2020M5 (May 2020) is 300%. *DE*
The approximate percentage increase in nominal expenditure (blue bars) from 2020M5 (May 2020) to 2020M9 (September 2020) is 25%. *DE*
The trend of nominal expenditure (blue bars) from 2020M1 to 2020M5 in the chart labeled 'Guatemala: Budgetary Arrears' is increasing. *PR*
The trend of Guatemala' s budgetary arrears (black line) from 2020M1 to 2020M5 in the chart labeled 'Guatemala: Budgetary Arrears' is increasing then decreasing. *PR*
The combination of trends in the chart that indicates Guatemala avoided fiscal collapse during the COVID-19 crisis is nominal expenditure (blue bars) increasing while arrears (black line) peak then decline. *LR*
The combination of trends in the chart that indicates Guatemala avoided fiscal collapse during the COVID-19 crisis is nominal expenditure (blue bars) increasing while arrears (black line) peak then decline. *LR*
The trend of nominal expenditure (blue bars) from 2020M5 to 2020M9 in the chart labeled 'Guatemala: Budgetary Arrears' is increasing. *LR*
The final answer is: \\boxed{A}. *DS*

Figure 18: Model performance before and after fine-tuning (Qwen2.5-VL-7B-Instruct)

Financial Markets & Products - Level: Decision Support - Degree: 5



Question:
When the Shanghai Index reached 2245.43, what is the most risk - averse strategy to protect against potential losses based on the chart's post - peak trend?

Options:
A: Sell all holdings immediately to secure profits before the market decline
B: Increase position size in high - growth sectors to capitalize on momentum
C: Hold existing positions and reinvest all dividends into new investments
D: Short the index to profit from the anticipated downward correction

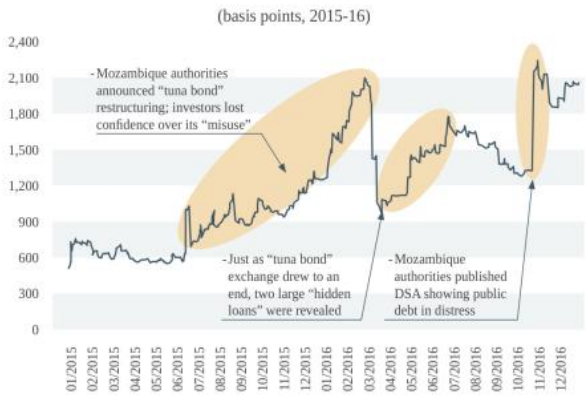
Image Background:
The image is labeled as 'Figure 70' and depicts a stock market chart referenced in the context of Ma Weipeng's investment experience. The subject is the Chinese stock market performance between late 1999 and 2001, specifically showing an upward trend of over 900 index points during the first year and a half of Ma Weipeng's participation in the market.

Qwen2.5-VL-7B-Instruct:
The question asks for the most risk-averse strategy to protect against potential losses when the Shanghai Index reached 2245.43, based on the chart's post-peak trend. The chart shows a peak at 2245.43, and the trend after this peak is downward, indicating a potential decline in the market.
Let's analyze each option:
- Option A: Sell all holdings immediately to secure profits before the market decline. This is a risk-averse strategy as it aims to lock in profits before the market drops.
- Option B: Increase position size in high-growth sectors to capitalize on momentum. This is a risky strategy as it involves increasing exposure to the market, which could lead to larger losses if the market declines.
- Option C: Hold existing positions and reinvest all dividends into new investments. This strategy involves staying in the market and reinvesting dividends, which could be risky if the market continues to decline.
- Option D: Short the index to profit from the anticipated downward correction. Shorting the index is a high-risk strategy as it involves borrowing assets to sell them, and if the market rises, the investor will lose money.
Given the post-peak trend and the potential for a market decline, the most risk-averse strategy would be to sell all holdings immediately to secure profits before the market decline. Final answer: \\boxed{A}

PyFi-QwenVL-7B:
The highest value of the Shanghai Index shown on the chart is 2245.43. *PP*
The lowest value of the Shanghai Index shown on the chart is 1047.83. *PP*
The approximate percentage increase of the Shanghai Index from its lowest point (1047.83) to its highest point (2245.43) is 115%. *DE*
The approximate percentage decrease of the Shanghai Index from its peak (2245.43) to its subsequent trough (1047.83) is 53%. *DE*
The Shanghai Index exhibits a peak followed by a sharp decline after reaching its highest point of 2245.43. *PR*
The Shanghai Index exhibits a sharp decline after reaching its peak of 2245.43, based on the chart. *PR*
Given the Shanghai Index's sharp decline after reaching 2245.43, the most risk - averse strategy to protect against potential losses is to sell all holdings immediately to secure profits before the market decline. *LR*
The Shanghai Index's sharp decline after reaching 2245.43 indicates that the market's momentum is exhausted, making it a suitable time to sell to avoid further losses. *LR*
When the Shanghai Index reached 2245.43, the most risk - averse strategy to protect against potential losses based on the chart's post - peak trend is to sell all holdings immediately to secure profits before the market decline. The final answer is: \\boxed{A}. *DS*

Figure 19: Model performance before and after fine-tuning (Qwen2.5-VL-7B-Instruct)

Bonds & Fixed Income - Level: Decision Support - Degree: 5



Question:

In January 2016, an investor who has seen Mozambique's sovereign bond spread rise due to transparency issues should most rationally decide to...

Options:

A: Purchase bonds immediately to capitalize on the dip in spreads after hidden loans were revealed

B: Avoid purchasing bonds due to the market's sensitivity to transparency issues, as seen by the sharp spread increases

C: Short sell bonds to profit from expected further spread decreases

D: Wait for the DSA report to confirm the extent of debt distress before investing

Image Background:

Figure 1.4 Lack of Transparency: Market Reactions. The image shows market reactions in Mozambique and Zambia due to lack of debt transparency. Abbreviations: WB (World Bank), IMF (International Monetary Fund), EMBI (Emerging Markets Bond Index). Acronyms and abbreviations in the image: DSA = Debt Sustainability Analysis;

Qwen2.5-VL-7B-Instruct:

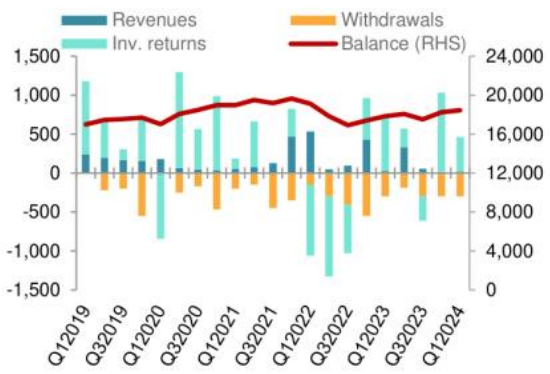
The image shows that in January 2016, there was a significant rise in Mozambique's sovereign bond spread due to transparency issues. The spread increased sharply, indicating a loss of investor confidence. Given this information, the most rational decision for an investor would be to avoid purchasing bonds due to the market's sensitivity to transparency issues, as seen by the sharp spread increases.
Option A suggests purchasing bonds immediately, which is not advisable given the current market conditions and the lack of transparency.
Option B advises avoiding bonds, which aligns with the current market sentiment and the sharp spread increases.
Option C suggests short selling bonds, which is speculative and not necessarily rational given the lack of transparency and the potential for further spread increases.
Option D suggests waiting for the DSA report, which might be too late as the spread has already increased significantly.
Therefore, the most rational decision is to avoid purchasing bonds due to the market's sensitivity to transparency issues.

PyFi-QwenVL-7B:

The unit of measurement for the bond spread values on the y-axis of the chart is basis points. *PP*
The approximate bond spread value (in basis points) for Mozambique in January 2016, as shown in the chart, was 1200. *DE*
The approximate bond spread value (in basis points) for Mozambique in July 2015, as shown in the chart, was 900. *DE*
The approximate percentage increase in Mozambique's sovereign bond spread from July 2015 to January 2016, based on the chart, is 33%. *DE*
The approximate percentage increase in Mozambique's sovereign bond spread from July 2015 to November 2016, based on the chart, is 100%. *DE*
The pattern observed in Mozambique's sovereign bond spread from July 2015 to November 2016, based on the chart, is that it increased sharply in July 2015, then decreased slightly, increased again in November 2016, and then decreased. *PR*
The pattern observed in Mozambique's sovereign bond spread from July 2015 to November 2016, based on the chart, is that it increased sharply in July 2015, then decreased slightly, increased again in November 2016, and then decreased. *PR*
The sharp increase in Mozambique's sovereign bond spread in July 2015 was caused by the revelation of hidden loans. *LR*
The sharp increase in Mozambique's sovereign bond spread in July 2015 was caused by the revelation of hidden loans. *LR*
In January 2016, an investor who has seen Mozambique's sovereign bond spread rise due to transparency issues should most rationally decide to avoid purchasing bonds due to the market's sensitivity to transparency issues, as seen by the sharp spread increases. *DS*
The final answer is: \\boxed{B}

Figure 20: Model performance before and after fine-tuning (Qwen2.5-VL-7B-Instruct)

Financial Markets & Products - Level: Decision Support - Degree: 5



Question:

Based on the Petroleum Fund's quarterly investment returns (light green bars) shown in the chart, which action would most effectively address the long-term threat of volatile investment returns to the Fund's balance stability?

Options:

A: Increase budget withdrawals to align with the Estimated Sustainable Income (ESI).

B: Develop strategies to stabilize and enhance investment returns.

C: Reduce petroleum revenue transfers to the Fund.

D: Lower the Estimated Sustainable Income (ESI) threshold.

Image Background:

The image is Figure 1.19 titled 'The petroleum fund balance stabilized for the first quarter this year partly due to contained withdrawals last year', showing the Petroleum Fund (PF) balance in USD million. The Petroleum Fund is referenced with the abbreviation PF. The image appears in the context of discussing the PF's capital value as of March 2024, which was USD 18.45 billion, and its 3.4 percent year-over-year growth from USD 17.83 billion. The source is INETL, World Bank Commodity Markets Reports, and World Bank Staff calculation.

Qwen2.5-VL-7B-Instruct:

The chart shows the balance of the Petroleum Fund (red line) and the investment returns (light green bars). The volatility in investment returns (light green bars) is a significant factor affecting the Fund's balance stability. To address the long-term threat of volatile investment returns, developing strategies to stabilize and enhance investment returns (option B) would be the most effective action. This would help in managing the Fund's balance more effectively over time. Increasing budget withdrawals to align with the Estimated Sustainable Income (ESI) (option A) might not directly address the volatility in investment returns. Reducing petroleum revenue transfers to the Fund (option C) could potentially reduce the Fund's balance but might not necessarily stabilize investment returns. Lowering the Estimated Sustainable Income (ESI) threshold (option D) could also indirectly affect the Fund's balance but does not directly address the volatility in investment returns.
Therefore, the most effective action to address the long-term threat of volatile investment returns is to develop strategies to stabilize and enhance investment returns.

PyFi-QwenVL-7B:

The color representing the Petroleum Fund's investment returns in the chart is light green. *PP*
The color representing the Petroleum Fund's revenues in the chart is Blue. *PP*
The color representing the Petroleum Fund's withdrawals in the chart is Orange. *PP*
The approximate value of the Petroleum Fund's investment returns (light green bars) in Q1 2020 is -1000 million USD. *DE*
The approximate value of the Petroleum Fund's investment returns (light green bars) in Q1 2024 is 500 million USD. *DE*
The approximate value of the Petroleum Fund's investment returns (light green bars) in Q3 2022 is -1000 million USD. *DE*
The approximate total value of the Petroleum Fund's investment returns (light green bars) from Q1 2020 to Q1 2024 is 1000 million USD. *DE*
The approximate total value of the Petroleum Fund's investment returns (light green bars) from Q1 2020 to Q1 2024 is 1000 million USD. *DE*
The pattern observed in the relationship between the Petroleum Fund's investment returns (light green bars) and its balance (red line) over the quarters shown is that investment returns are volatile, and the balance is relatively stable. *PR*
The volatility of the Petroleum Fund's investment returns (light green bars) is most likely caused by the Fund's investment strategy. *LR*
The volatility of the Petroleum Fund's investment returns (light green bars) is most likely caused by the Fund's investment strategy. *LR*
The action that would most effectively address the volatility of the Petroleum Fund's investment returns (light green bars) shown in the chart is to develop strategies to stabilize and enhance investment returns. The final answer is: \\boxed{B} *DS*

Figure 21: Model performance before and after fine-tuning (Qwen2.5-VL-7B-Instruct)

H Detailed Results

H.1 Experiments Consideration and Ablation Studies

In order to comprehensively evaluate the performance of various VLMs on our samples from ***PyFi-600K*** and eliminate potential biases, we consider the following factors and conduct comparative experiments and ablation studies accordingly.

The Impact of Contextual Information. To investigate the impact of image-related contextual information on model performance, we provide the target model with an image and a question, followed by additional textual information in one of four conditions: (1) providing image background information only, (2) providing image contextual analysis only, (3) providing both background information and contextual analysis, and (4) providing no additional text. The background information includes the image's title, topic, and the full forms of all abbreviations, offering necessary context for answering the question without revealing the answer itself. The contextual analysis, generated by a large language model, summarizes relevant contextual information present in the image and offers direct guidance for solving the problem. This experimental design aims to determine whether the model genuinely understands the visual content or relies on external textual information to infer the answer.

Risk of Data Leakage. The VLMs evaluated in this study were pretrained on massive corpora. Since our data primarily consists of publicly available financial documents sourced from the internet and predominantly dated before 2025, there is a potential risk of data leakage. To address this concern, we additionally evaluated the VLMs in a setting where images were not provided. By comparing model accuracy between the two settings, with and without images, we can assess whether the pretraining corpus might already contain images or questions from our dataset. As for the formal statistics, we exclude those samples that can be correctly answered by more than half of the evaluated models even without image input, and the remaining samples are defined as Evaluation Dataset after Exclusion while the initial ones are defined as Original Evaluation Dataset.

Impact of Option Distribution. Our analysis of the ground-truth labels reveals severe class imbalance. Specifically, 43.39% of the samples designate option A as the correct answer, 34.20% select option B, 15.81% choose option C, and only 6.08% pick option D. All remaining options combined account for less than 1% of the total answers. Details can be found in the appendix 7. Such extreme imbalance may introduce bias during both training and evaluation, potentially undermining model performance. To address this issue and obtain more robust and reliable evaluation results, we rebalance the sample options so that, in a set of one thousand samples, each of the four options A, B, C, and D constitutes approximately 25%. We then compare model performance before and after this rebalancing.

All experimental results are presented in the following sections. The meanings of key abbreviations and the group configurations are defined in the tables below.

Table 9: Meaning of Abbreviations

Abbreviation	Meaning
PP	Perception
DE	Data Extraction
CA	Calculation Analysis
PR	Pattern Recognition
LR	Logical Reasoning
DS	Decision Support

Table 10: Experimental Group Configurations

Group	Input Configuration
I	Image and Image Background
II	Image and Analysis Information
III	Image, Image Background and Analysis Information
IV	Image Only

H.2 Detailed Experiment Results on Original Evaluation Dataset with Image

Table 11 presents the overall accuracy and performance across groups for all models on the original evaluation dataset (1000 samples). GLM-4.5V achieves the highest overall accuracy (85.22%) among all pre-trained models, significantly outperforming others. Claude-opus follows as the second-best pre-trained model (79.88%). Among the Qwen series, the larger models generally perform better, with Qwen3-VL-Plus (76.70%) leading, showing a clear scaling effect from 3B (59.00%) to 72B (74.28%). For SFT models, fine-tuning with 500 or 47K samples shows modest but inconsistent improvements over their base Instruct-L versions. For instance, PyFi-QwenVL-3B-COT-47K reaches

Table 11: Model Performance Comparison By Group On Original Evaluation Dataset (Samples=1000, Groups with Image)

Model	Overall	TC	TQ	I	II	III	IV
Pre-trained VLMs							
GPT-4.1	76.08	3043	4000	75.4	75.0	77.5	76.4
InternVL3-38B	73.72	2949	4000	73.9	74.6	73.7	72.7
Claude-opus-4-1-20250805	79.88	3195	4000	80.8	79.5	80.6	78.6
GLM-4.5V	85.22	3409	4000	85.8	85.3	85.5	84.3
DeepSeek-VL2	63.20	2528	4000	61.8	63.3	65.6	62.1
Hunyuan-Large-Vision	76.60	3064	4000	77.3	76.4	77.5	75.2
ERNIE-4.5-turbo-vl	68.83	2753	4000	68.0	70.0	71.1	66.2
Moonshot-V1-128k-Vision-Preview	76.40	3056	4000	75.3	77.4	78.2	74.7
Moonshot-V1-32k-Vision-Preview	76.22	3049	4000	75.1	77.1	77.8	74.9
Moonshot-V1-8k-Vision-Preview	76.53	3061	4000	75.0	77.5	77.8	75.8
Qwen3-VL-Plus	76.70	3068	4000	76.5	76.6	76.8	76.9
Qwen2.5-VL-72B-Instruct	74.28	2971	4000	72.9	75.4	75.5	73.3
Qwen2.5-VL-32B-Instruct	71.78	2871	4000	71.2	72.4	73.0	70.5
Qwen2.5-VL-7B-Instruct	67.55	2702	4000	65.0	68.4	69.7	67.1
Qwen2.5-VL-3B-Instruct	59.00	2360	4000	57.3	59.7	62.1	56.9
SFT VLMs, Trained with 500 Sample Chains							
Qwen2.5-VL-3B-Instruct-L	58.80	2352	4000	58.4	60.3	60.0	56.5
PyFi-QwenVL-3B-500	61.52	2461	4000	60.6	62.8	63.6	59.1
PyFi-QwenVL-3B-COT-500	62.32	2493	4000	61.8	63.9	63.3	60.3
Qwen2.5-VL-7B-Instruct-L	66.60	2664	4000	65.5	67.5	68.4	65.0
PyFi-QwenVL-7B-500	66.27	2651	4000	62.3	67.1	67.2	68.5
PyFi-QwenVL-7B-COT-500	65.72	2629	4000	64.4	66.4	66.0	66.1
SFT VLMs, Trained with 47K Sample Chains							
Qwen2.5-VL-3B-Instruct-L	58.80	2352	4000	58.4	60.3	60.0	56.5
PyFi-QwenVL-3B-47K	61.85	2474	4000	61.6	63.0	62.5	60.3
PyFi-QwenVL-3B-COT-47K	64.40	2576	4000	64.2	63.6	66.6	63.2
Qwen2.5-VL-7B-Instruct-L	66.60	2664	4000	65.5	67.5	68.4	65.0
PyFi-QwenVL-7B-47K	66.35	2654	4000	65.7	67.1	67.2	65.4
PyFi-QwenVL-7B-COT-47K	67.42	2697	4000	67.8	67.6	68.0	66.3

64.40%, a noticeable gain over the base Qwen2.5-VL-3B-Instruct (59.00%). However, for the 7B models, gains from SFT are marginal, with some variants even underperforming the base model. This indicates that the effectiveness of SFT may be model-size and data-dependent.

The following tables present performance on the Evaluation Dataset after Exclusion, which removes samples that more than half the models could answer correctly without the image. This subset (301 samples) is designed to mitigate data leakage concerns and focus on problems that genuinely require visual understanding.

Table 12 provides an overall summary on the excluded dataset, including total correct (TC) out of total questions (TQ = 4 groups * 301 samples = 1204). The ranking of pre-trained models remains consistent. It also shows performance per group (I, II, III, IV). The top models like GLM-4.5V and Claude-opus maintain high and consistent accuracy across all four groups, demonstrating robustness. The SFT results reinforce previous observations: The 47K COT training for the 3B model (PyFi-QwenVL-3B-COT-47K) yields the most dramatic improvement, boosting overall accuracy from 20.85% to 40.37% and showing gains across all groups. For the 7B models, the 500-sample SFT (PyFi-QwenVL-7B-500) provides a strong boost, but the 47K COT variant also delivers competitive performance. Notably, the non-COT 47K SFT for the 7B model (PyFi-QwenVL-7B-47K) underperforms significantly, highlighting the critical importance of the Chain-of-Thought

Table 12: Model Performance Comparison By Group On Evaluation Dataset After Exclusion (Samples=301, Groups with Image)

Model	Overall	TC	TQ	I	II	III	IV
Pre-trained VLMs							
GPT-4.1	52.99	638	1204	51.5	52.16	52.82	55.48
InternVL3-38B	52.91	637	1204	55.81	52.16	50.17	53.49
Claude-opus-4-1-20250805	64.70	779	1204	65.78	62.46	63.12	67.44
GLM-4.5V	74.75	900	1204	74.75	75.08	73.75	75.42
DeepSeek-VL2	45.18	544	1204	45.85	42.86	45.18	46.84
Hunyuan-Large-Vision	59.72	719	1204	60.13	60.8	59.47	58.47
ERNIE-4.5-turbo-vl	34.47	415	1204	32.56	34.55	34.88	35.88
Moonshot-V1-128k-Vision-Preview	54.57	657	1204	53.82	54.15	54.15	56.15
Moonshot-V1-32k-Vision-Preview	54.40	655	1204	54.82	53.82	53.16	55.81
Moonshot-V1-8k-Vision-Preview	54.90	661	1204	53.82	54.15	54.15	57.48
Qwen3-VL-Plus	51.00	614	1204	52.16	49.17	48.17	54.49
Qwen2.5-VL-72B-Instruct	48.84	588	1204	47.18	47.51	46.51	54.15
Qwen2.5-VL-32B-Instruct	43.19	520	1204	40.86	42.19	42.19	47.51
Qwen2.5-VL-7B-Instruct	37.87	456	1204	36.21	36.21	36.54	42.52
Qwen2.5-VL-3B-Instruct	20.51	247	1204	18.6	18.94	21.26	23.26
SFT VLMs, Trained with 500 Sample Chains							
Qwen2.5-VL-3B-Instruct-L	20.85	251	1204	20.6	21.93	17.28	23.59
PyFi-QwenVL-3B-500	26.74	322	1204	28.57	25.25	23.92	29.24
PyFi-QwenVL-3B-COT-500	27.66	333	1204	29.57	25.58	21.59	33.89
Qwen2.5-VL-7B-Instruct-L	34.55	416	1204	34.55	32.89	32.23	38.54
PyFi-QwenVL-7B-500	43.44	523	1204	39.53	41.53	40.2	52.49
PyFi-QwenVL-7B-COT-500	39.53	476	1204	40.86	38.87	34.88	43.52
SFT VLMs, Trained with 47K Sample Chains							
Qwen2.5-VL-3B-Instruct-L	20.85	251	1204	20.6	21.93	17.28	23.59
PyFi-QwenVL-3B-47K	25.25	304	1204	25.58	24.25	22.92	28.24
PyFi-QwenVL-3B-COT-47K	40.37	486	1204	40.2	37.21	38.87	45.18
Qwen2.5-VL-7B-Instruct-L	34.55	416	1204	34.55	32.89	32.23	38.54
PyFi-QwenVL-7B-47K	27.08	326	1204	26.58	25.91	24.92	30.9
PyFi-QwenVL-7B-COT-47K	42.61	513	1204	44.19	38.87	42.52	44.85

training methodology alongside data volume for achieving gains on this challenging visual reasoning benchmark.

Table 13 breaks down model performance by required capability on the challenging 301-sample subset. The overall accuracy drops dramatically for all models compared to the full dataset, confirming the difficulty of these leakage-filtered samples. GLM-4.5V again dominates with 74.75% overall, excelling particularly in Perception (PP, 89.47%) and Data Extraction (DE, 78.07%). Claude-opus (64.70%) shows strong performance in Calculation Analysis (CA, 66.02%) and Logical Reasoning (LR, 72.58%). Hunyuan-Large-Vision is notable for its high Perception (90.79%) and relatively strong Decision Support (DS, 51.92%). A clear capability-wise difficulty trend emerges: Perception is generally easiest for top models, while Calculation Analysis and, especially, Decision Support are most challenging. Among SFT models, PyFi-QwenVL-3B-COT-47K shows a remarkable leap to 40.37%, more than doubling the performance of its base model (20.85%) and significantly outperforming its 500-sample counterpart. This suggests that Chain-of-Thought (COT) training on a larger, high-quality dataset (47K) is highly effective for smaller models, particularly improving Pattern Recognition (PR) and Logical Reasoning (LR). For the 7B models, PyFi-QwenVL-7B-500 achieves the best overall (43.44%), but the 47K COT variant also shows strong gains in LR.

Table 14 analyzes performance stratified by

Table 13: Model Performance Comparison By Capability On Evaluation Dataset After Exclusion (Samples=301, Groups with Image)

Model	Overall	PP	DE	CA	PR	LR	DS
Pre-trained VLMs							
GPT-4.1	52.99	86.84	47.64	40.62	57.14	54.84	38.46
InternVL3-38B	52.91	76.32	56.37	44.53	43.88	53.23	30.77
Claude-opus-4-1-20250805	64.70	80.92	62.74	66.02	58.16	72.58	32.69
GLM-4.5V	74.75	89.47	78.07	65.23	75.51	75.81	46.15
DeepSeek-VL2	45.18	76.32	44.58	36.33	42.35	42.74	19.23
Hunyuan-Large-Vision	59.72	90.79	58.96	36.72	63.27	69.35	51.92
ERNIE-4.5-turbo-vl	34.47	34.87	33.96	36.72	25.00	45.16	36.54
Moonshot-V1-128k-Vision-Preview	54.57	80.92	61.08	39.06	41.84	58.87	38.46
Moonshot-V1-32k-Vision-Preview	54.40	80.26	61.56	37.11	40.31	62.90	38.46
Moonshot-V1-8k-Vision-Preview	54.90	80.26	61.79	39.84	42.35	58.06	38.46
Qwen3-VL-Plus	51.00	77.63	54.25	37.50	41.84	57.26	32.69
Qwen2.5-VL-72B-Instruct	48.84	70.39	59.91	33.20	35.71	46.77	26.92
Qwen2.5-VL-32B-Instruct	43.19	60.53	45.28	37.50	36.73	42.74	28.85
Qwen2.5-VL-7B-Instruct	37.87	54.61	49.53	34.77	18.37	24.19	15.38
Qwen2.5-VL-3B-Instruct	20.51	36.84	15.57	21.48	16.84	21.77	19.23
SFT VLMs, Trained with 500 Sample Chains							
Qwen2.5-VL-3B-Instruct-L	20.85	35.53	17.22	21.09	15.82	20.97	25.00
PyFi-QwenVL-3B-500	26.74	34.87	28.77	28.52	16.84	24.19	21.15
PyFi-QwenVL-3B-COT-500	27.66	37.50	28.77	26.17	22.45	25.81	21.15
Qwen2.5-VL-7B-Instruct-L	34.55	52.63	42.92	31.64	18.37	25.00	11.54
PyFi-QwenVL-7B-500	43.44	71.05	44.58	43.36	30.10	37.90	17.31
PyFi-QwenVL-7B-COT-500	39.53	61.84	41.51	40.23	26.02	34.68	17.31
SFT VLMs, Trained with 47K Sample Chains							
Qwen2.5-VL-3B-Instruct-L	20.85	35.53	17.22	21.09	15.82	20.97	25.00
PyFi-QwenVL-3B-47K	25.25	42.76	19.81	18.75	29.59	29.03	25.00
PyFi-QwenVL-3B-COT-47K	40.37	62.50	36.56	29.69	41.84	48.39	34.62
Qwen2.5-VL-7B-Instruct-L	34.55	52.63	42.92	31.64	18.37	25.00	11.54
PyFi-QwenVL-7B-47K	27.08	36.84	23.35	30.47	17.35	37.10	25.00
PyFi-QwenVL-7B-COT-47K	42.61	68.42	41.75	34.77	32.65	49.19	34.62

problem complexity degree (1=Least, 5=Most Complex). A strong negative correlation between model performance and complexity degree is evident for most models. For example, GLM-4.5V scores 87.75% on degree-1 problems but drops to 64.22% on degree-4. Interestingly, performance on degree-5 (the most complex) does not always follow this monotonic decline; for models like Claude-opus, Hunyuan, and some Qwen variants, degree-5 accuracy is higher than degree-4. This may indicate that the highest complexity problems, while difficult, sometimes involve different (e.g., more structured) reasoning that some models can handle. The SFT models show interesting patterns: PyFi-QwenVL-3B-COT-47K demonstrates substantial improvements across all complexity levels, most notably on degrees 4 and 5 (40.09% and 48.08%, up from 21% and 29% in the base model). This suggests that the COT training on extensive data particularly enhances the model's ability to tackle high-complexity, multi-step reasoning problems. The PyFi-QwenVL-7B-COT-47K model also shows strong gains on degree-5 problems (59.62%).

Summary of Key Findings

Performance Gap. A significant performance gap exists between top-tier proprietary models (GLM-4.5V, Claude-opus) and open-source models, especially on the challenging leakage-filtered subset.

Scaling Laws. For the Qwen family, model size strongly correlates with performance on this financial visual QA task.

Effectiveness of SFT: Supervised Fine-Tuning can

Table 14: Model Performance Comparison By Complexity Degree On Evaluation Dataset After Exclusion (Samples=301, Groups with Image)

Model	Overall	1	2	3	4	5
Pre-trained VLMs						
GPT-4.1	52.99	74.51	46.12	52.63	46.55	53.85
InternVL3-38B	52.91	75.00	52.91	41.78	44.83	67.31
Claude-opus-4-1-20250805	64.70	76.47	61.89	66.12	55.60	73.08
GLM-4.5V	74.75	87.75	76.21	72.37	64.22	73.08
DeepSeek-VL2	45.18	67.65	43.45	34.54	42.24	46.15
Hunyuan-Large-Vision	59.72	80.88	54.85	50.00	57.76	80.77
ERNIE-4.5-turbo-vl	34.47	32.84	33.25	38.82	25.43	65.38
Moonshot-V1-128k-Vision-Preview	54.57	74.02	56.31	46.71	43.97	57.69
Moonshot-V1-32k-Vision-Preview	54.40	73.53	56.31	45.72	43.97	61.54
Moonshot-V1-8k-Vision-Preview	54.90	73.53	56.80	48.03	43.97	55.77
Qwen3-VL-Plus	51.00	72.55	50.00	45.72	37.50	65.38
Qwen2.5-VL-72B-Instruct	48.84	67.16	57.77	36.51	31.03	57.69
Qwen2.5-VL-32B-Instruct	43.19	58.33	43.69	39.14	30.60	59.62
Qwen2.5-VL-7B-Instruct	37.87	55.39	44.90	31.91	17.67	38.46
Qwen2.5-VL-3B-Instruct	20.51	31.86	12.62	23.03	18.97	30.77
SFT VLMs, Trained with 500 Sample Chains						
Qwen2.5-VL-3B-Instruct-L	20.85	33.33	13.11	21.38	21.12	28.85
PyFi-QwenVL-3B	26.74	40.69	20.87	28.29	20.69	36.54
PyFi-QwenVL-3B-COT	27.66	39.22	22.57	27.96	24.57	34.62
Qwen2.5-VL-7B-Instruct-L	34.55	55.88	37.86	30.26	15.52	34.62
PyFi-QwenVL-7B	43.44	66.18	44.17	39.47	28.45	38.46
PyFi-QwenVL-7B-COT	39.53	58.82	39.32	36.51	23.71	53.85
SFT VLMs, Trained with 47K Sample Chains						
Qwen2.5-VL-3B-Instruct-L	20.85	33.33	13.11	21.38	21.12	28.85
PyFi-QwenVL-3B	25.25	37.25	18.93	19.74	31.47	32.69
PyFi-QwenVL-3B-COT	40.37	59.80	32.77	36.51	40.09	48.08
Qwen2.5-VL-7B-Instruct-L	34.55	55.88	37.86	30.26	15.52	34.62
PyFi-QwenVL-7B	27.08	35.29	21.60	26.64	25.86	46.15
PyFi-QwenVL-7B-COT	42.61	62.25	38.35	39.80	32.76	59.62

substantially improve smaller models (3B), particularly when using a large (47K), high-quality dataset with Chain-of-Thought reasoning steps. Gains for larger models (7B) are more modest and sensitive to the training recipe.

Capability & Complexity. Perception and Data Extraction are relative strengths for leading models, while Calculation Analysis and Decision Support remain challenging. Problem complexity is inversely correlated with accuracy, but advanced training (COT) can specifically improve performance on high-complexity problems.

I Algorithms

The proposed adversarial agent dataset construction method is outlined in Algorithm 1. Its runtime is primarily dictated by the base VLM's inference latency. Our implementation, using Doubao-seed-1-6-flash-250828, completes approximately 100 adversarial rounds (100 samples) in about 60 minutes. While the MCTS process for a single image runs sequentially, each adversarial step can generate multiple branches in parallel for internal acceleration. The framework also enables concurrent processing of multiple images across servers and processes. Consequently, the approach is readily extensible to large-scale synthesis with manageable time costs.

Algorithm 1: Dataset Construction via Adversarial Agents under the MCTS Paradigm

Input: Challenger Agent ψ, Solver Agent ϕ, Image I with background b_I and analysis info f_I, capability levels $\mathcal{L} = \{l_1, \ldots, l_6\}$, complexity levels $\mathcal{C} = \{c_1, \ldots, c_5\}$, MCTS parameters α, β, max chains $T_{\max}$

Output: Pyramid-structured dataset $\mathcal{S}$ with question-answer chains

Generate final question Q_F using LLMs with: b_I, f_I, $l(Q_F) = l_6$,
Determine ground truth answer A_F for Q_F via multi-LLM consensus
Initialize MCTS tree $\mathcal{T}$ with root Q_F
$t \leftarrow 0$
while $t < T_{\max}$ *and* $\mathcal{T}$ *not terminated* **do**
  $\psi, \phi \leftarrow$ root of $\mathcal{T}$
  while ψ *and* ϕ *ready and debate not terminated* **do**
    Compete (Challenger Agent ψ Turn)
    $N_{\text{child}} \leftarrow$ children of q_ψ
    if $|N_{child}| = 0$
      Determine current and next capability level l_{curr}, l_{next} based on pyramid progression
      Generate new question q_{new} with: b_I, $l_{\text{curr}} = l_{\text{next}}$, complexity $\in \mathcal{C}$, visit_count $= 1$, victory_count $= 0$
      $q_\psi \leftarrow q_{\text{new}}$
    else
      $p_{\text{gen}} \leftarrow \sigma(\alpha \cdot (|N_{\text{child}}| - \beta))$
      Sample $z \sim \text{Bernoulli}(p_{\text{gen}})$
      if $z = 1$
        Determine current and next capability level l_{curr}, l_{next} based on pyramid progression
        Generate new question q_{new} with: b_I, capability $= l_{\text{next}}$, complexity $\in \mathcal{C}$, visit_count $= 1$, victory_count $= 0$
        $q_\psi \leftarrow q_{\text{new}}$
      else
        $q_\psi \leftarrow \arg\max_{n' \in N_{\text{child}}} \text{UCT}(n')$
        Increment visit_count(q_ψ)
      end
    end
    Compete (Solver Agent ϕ Turn)
    Generate answer a_ϕ to q_ψ
    if $a_\phi \in$ *existing actions*
      Increment visit_count(a_ϕ)
    else
      Create new a_ϕ with visit_count $\leftarrow 1$, victory_count $\leftarrow 0$
    end
    Add edge (q_ψ, a_ϕ) to $\mathcal{T}$
    if l_{curr} *= 6 and LLM judges chain can answer* Q_F
      Generate final answer a_{final}
      if $a_{final} = A_F$
        Backpropagate and increment victory_count for all nodes
      end
      Break
    end
  end
  $t \leftarrow t + 1$
end
Add successful chains to $\mathcal{S}$