FinMaster: A Holistic Benchmark for Full-Pipeline Financial Management with Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Financial management tasks are pivotal to global economic stability; however, their efficient execution faces persistent challenges, including labor intensive processes, low error tolerance, data fragmentation, and limitations in existing technological tools. Although large language models (LLMs) have shown remarkable success in various natural language processing (NLP) tasks and have demonstrated potential in automating workflows through reasoning and contextual understanding, current benchmarks for evaluating LLMs in finance suffer from insufficient domain-specific data, simplistic task design, and incomplete evaluation frameworks. To address these gaps, in this work, we present **FinMaster**, a comprehensive financial management benchmark designed to systematically assess the capabilities of LLM in financial literacy, accounting, auditing, and consulting. Specifically, **FinMaster** comprises three main modules: i) *FinSim*, which builds simulators that can generate synthetic, privacy-compliant financial datasets for different types of companies to replicate real-world market dynamics; ii) *FinSuite*, which provides a variety of tasks in core financial domains, spanning 183 tasks of various types and difficulty levels; and iii) *FinEval*, which develops a unified evaluation framework for streamlined evaluation. Extensive experiments on state-of-the-art LLMs, such as GPT-4o-mini, Claude-3.7-Sonnet, and DeepSeek-V3, reveal critical capability gaps in financial reasoning, with accuracy dropping from over 90% on basic tasks to merely 40% on complex scenarios requiring multi-step reasoning. This degradation exhibits the propagation of computational errors, where single-metric calculations that initially demonstrated 58% accuracy decreased to 37% in multimetric scenarios. To the best of our knowledge, **FinMaster** is the first benchmark that comprehensively covers full-pipeline financial workflows with challenging and realistic tasks. We hope that **FinMaster** can bridge the gap between the research community and industry practitioners, driving the adoption of LLMs in real-world financial practices to enhance both efficiency and accuracy.

1 Introduction

Financial management serves as the cornerstone of the global economic system, where financial tasks are critical for ensuring accuracy, compliance, and efficiency in economic activities, e.g., capital allocation, risk management, and investment strategic decision-making. The global financial services market reached $25.8 trillion in 2022 and is projected to grow to $37 trillion by 2027, with an annual growth rate of 7.4% (ReportLinker, 2023). This scale underscores the critical importance of efficient financial management systems and the potential impact of technological innovations in this domain. However, their execution faces challenges: **(i) Labor-intensive processes:** traditional financial tasks, such as accounting and auditing, rely heavily on manual operations, and financial professionals require extensive training to master complex regulations, which are repetitive and time-consuming; **(ii) Low error tolerance:** minor mistakes in financial statements, e.g., decimal errors or misclassifications, can trigger compliance risks or market volatility; **(iii) Data fragmentation:** financial data originates from diverse sources, each with unique data structures and update frequencies. However, the latency in real-time scenarios and the poor compatibility between systems in integrating heterogeneous data may lead to data silos; **(iv) Tool limitations:** existing fintech tools and rule-based systems often exhibit limitations in interpreting implicit logic in financial data and

adapting to evolving regulations, making it difficult for them to handle complex causal reasoning tasks. The advancements in LLMs, e.g., GPT-4 (Achiam et al., 2023) and DeepSeek-V3 (Liu et al., 2024), have demonstrated remarkable successes in tasks requiring reasoning, contextual understanding, and multi-step problem-solving across domains, e.g., code generation and math reasoning (Joel et al., 2024; Satpute et al., 2024). Their general-purpose capabilities make LLMs well-suited for automating financial management.

Several recent attempts demonstrate the potential of applying LLMs to finance tasks. FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) introduce complex numerical reasoning over financial statements but are constrained by static data and structured formats. FinBen (Xie et al., 2024) offers broader evaluation across diverse financial tasks but lacks granularity for domain-specific reasoning nuances. FinTSB (Hu et al., 2025) specializes in time-series forecasting but neglects the external factors. Other benchmarks like PIXIU (Xie et al., 2023), FinanceBench (Islam et al., 2023), and BizBench (Koncel-Kedziorski et al., 2023) exhibit a narrow scope, primarily focusing on conventional financial NLP tasks while overlooking complex financial reasoning and real-world applications. SECQUE (Yoash et al., 2025) advances LLMs evaluation by focusing on practical financial tasks requiring multi-step reasoning, but still relies on a static dataset and fails to fully capture the reality of financial management. DOCMATH-EVAL and FinanceMATH (Zhao et al., 2024b;a) also focus on the advanced reasoning in financial statements. Critically, existing benchmarks focus exclusively on understanding and reasoning over pre-prepared financial statements, neglecting the generation and auditing processes that transform raw transaction data into these statements. This represents a fundamental disconnect from real-world financial management. Without incorporating the full workflow, LLM evaluations remain artificially simplified and fail to address the real professional financial management.

	Fin Statement Tasks			Inf	Holistic
	Generate	Audit	Analyze	data	eval
FinQA	✗	✗	✓	✗	✗
PIXIU	✗	✗	✓	✗	✗
FinanceBench	✗	✗	✓	✗	✗
FinBen	✗	✗	✓	✗	✓
FinEval	✗	✗	✓	✗	✗
SECQUE	✗	✗	✓	✗	✗
FinanceMath	✗	✗	✓	✗	✓
FinMaster	✓	✓	✓	✓	✓

Table 1: Comparison of financial benchmarks.

To address these issues, we present **FinMaster**, a holistic benchmark for mastering full-pipeline financial management with LLMs. Specifically, **FinMaster** has three main modules: i) *FinSim*, a simulator that can automatically generate financial data, including transaction records and financial statements, for financial literacy, accounting, auditing and consulting. It can produce synthetic datasets for various types of companies to replicate real-world market dynamics, addressing the issue of lack of financial data due to privacy constraints. ii) FinSuite, a task suite comprising 183 tasks across accounting, auditing, and consulting domains, designed to evaluate LLM capabilities at varying difficulty levels. iii) *FinEval*, a unified evaluation framework for systematic evaluation of LLMs for quantifying the performances of state-of-the-art LLMs across different tasks with in-depth analysis such as accuracy and token usage. Extensive experiments over widely-used LLMs, e.g., GPT-4o-mini, Claude-3.7-Sonnet, DeepSeek-V3, and OpenAI o3-mini, show systematic gaps in financial reasoning capabilities. While these models achieve over 90% accuracy on basic tasks, their performance drops sharply to 40% in complex multi-step reasoning scenarios. Moreover, general-purpose LLMs lack domain-specific knowledge, leading to hallucinated conclusions or statistically invalid outputs when professional judgment is needed. To the best of our knowledge, **FinMaster** is the first financial benchmark that simulates multi-step financial operations for LLMs, serving as the fundamental testbed for the advanced LLMs. The code is released at `https://anonymous.4open.science/r/Finmaster-3957`.

2 Preliminaries

Transactions and Financial Statements. Transactions refer to the economic events or business activities that result in measurable changes in a company's financial position (Westermeier, 2020). These events are systematically recorded to populate accurate, reliable financial statements. There are three key financial statements that form the bedrock of financial reporting and analysis: i) income statement, which dynamically reports a company's revenues, expenses, and profits/losses over a specific period (SHARE, 1995); ii) balance sheet, which provides a snapshot of a company's financial position at a specific moment; and iii) cash flow

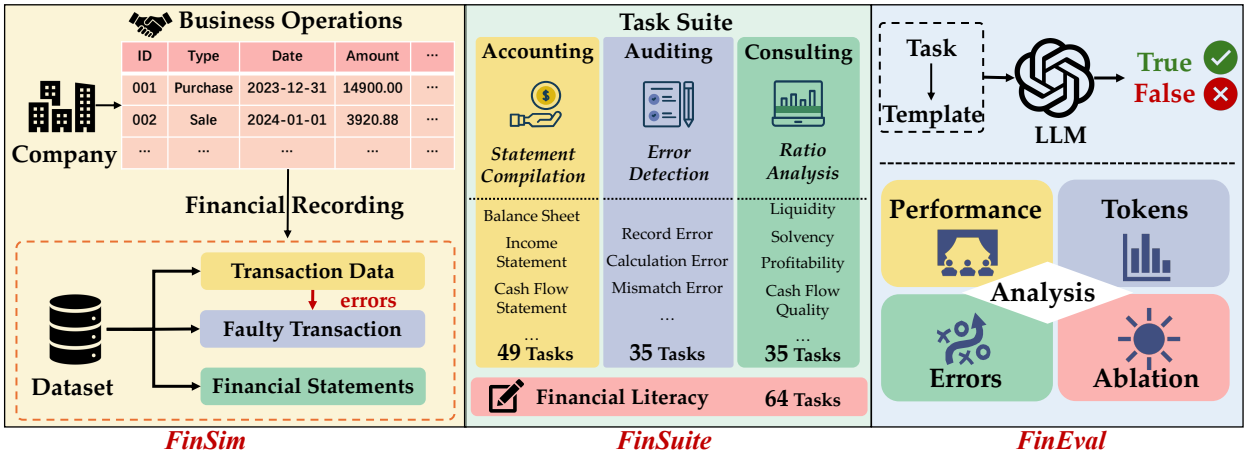


Figure 1: The three main modules of **FinMaster**.

statement, which tracks the actual cash movements, i.e., cash inflows and outflows, over a period. These financial statements are intrinsically interconnected, where the net income from the income statement flows into the retained earnings on the balance sheet, and the changes in the balance sheet affect the cash flow statement (White et al., 2002). Financial statements must follow standardized guidelines, e.g., Generally Accepted Accounting Principles (GAAP) (Epstein et al., 2009), to ensure consistency, transparency, and comparability, providing a comprehensive view of a company's financial performance.

Financial Management. **Accounting** is the systematic recording, analysis, and reporting of financial transactions to ensure transparency and regulatory compliance (Godfrey et al., 2010). It supports business operations by transforming raw transaction data into structured financial reports that inform strategic business decisions. Daily operations are recorded, classified, and adjusted to align with accounting principles, e.g., matching income and expenses, which form the basis for financial statements. **Auditing** is an independent assurance activity that verifies the compliance of a company's economic activities and the reliability of its financial information through systematic review and professional evaluation. Its core objective is to ensure the accuracy and completeness of financial data, with the verification of underlying transaction data serving as the foundation of the audit process. **Consulting** provides expert analysis to improve business performance (Biech, 2019; Bruhn et al., 2018). Financial diagnostics, i.e., evaluating profitability, operational efficiency, and solvency via frameworks like DuPont Analysis (Soliman, 2008) and Altman Z-scores (Altman et al., 2017), identify inefficiencies and competitive positioning. This analysis bridges financial data to strategic decisions, distinguishing consultants as data-driven problem solvers.

3 FinMaster

3.1 *FinSim*: Financial Data Simulator

We develop FinSim, a financial data simulator that models diverse company archetypes and generates comprehensive financial datasets including transaction records and financial statements.

	Type I	Type II	Type III	Type IV	Type V
Initial Capital	28M	13M	13M	13M	16M
Fixed Asset Purchase Freq	0.00/2.00	1.00/2.00	1.00/2.00	0.00/1.00	0.00/2.00
Purchase Unit Price	950,000	45,000	21,250	31,500	1,823
Profit Margin	0.30/0.50	0.10/0.40	0.70/1.00	0.80/2.00	0.30/0.80
Quantity Per Purchase	1.00	15.00	5.00	2.00	500.00
Purchase Frequency	1.00/2.00	1.00/3.00	2.00/4.00	0.00/2.00	1.00/3.00
Credit Purchase Ratio	0.1	0.1	0.3	0.3	0.6
Quantity Per Sale	1.00	5.00	3.00	1.00	5.00
Sales Frequency	0.00/1.00	1.00/2.00	2.00/4.00	0.00/3.00	2.00/4.00
Credit Sales Ratio	0.6	0.4	0.3	0.7	0.4
Expense Frequency	1.00/2.00	2.00/4.00	2.00/3.00	1.00/2.00	1.00/2.00

Table 2: *FinSim* configurations for company types.

Types of Companies. To reflect real-world market diversity, *FinSim* incorporates five company archetypes from audited financial statements: Type I (capital goods manufacturers with intensive operations, substantial fixed assets, infrequent high-value transactions); Type II (transaction-driven enterprises with standardized costs, limited pricing power, volume-

dependent models); Type III (high value-added consumer goods with superior margins, efficient production, brand investments); Type IV (asset-light service providers with minimal fixed assets, exceptional margins); and Type V (high-turnover retail with frequent low-price transactions, substantial volumes). Table 2 presents detailed financial parameters across diverse business models.

Types of Transactions. Several types of transactions or financial records are considered in *FinSim* to simulate the complexity of realistic operations while maintaining the integrity of financial systems. i) Asset data: including the initializations of cash deposit, bank deposit, and fixed assets. ii) Operational data: including purchase management, which refers to the purchase-related transactions with different payment methods, e.g., cash and bank transfer; sales management, which refers to the sales-related transactions, corresponding to the purchase-related transactions; and fixed assets management, which comprises fixed asset purchase transactions, i.e., recording the fixed asset acquisitions, and fixed asset depreciation, i.e., generating monthly depreciation entries on the 1st of each month. iii) Financial data: including cash flow management, which ensures the cash balance, triggering cash-bank transfers, i.e., bank to cash transfer transactions and cash to bank transfer transactions, when cash balances fall below a threshold; expense processing, which logs different types of expenses, including administrative expenses, sales expenses, and financial expenses; and some other transactions such as interest receivables, which create interest receivable entries for bank deposits.

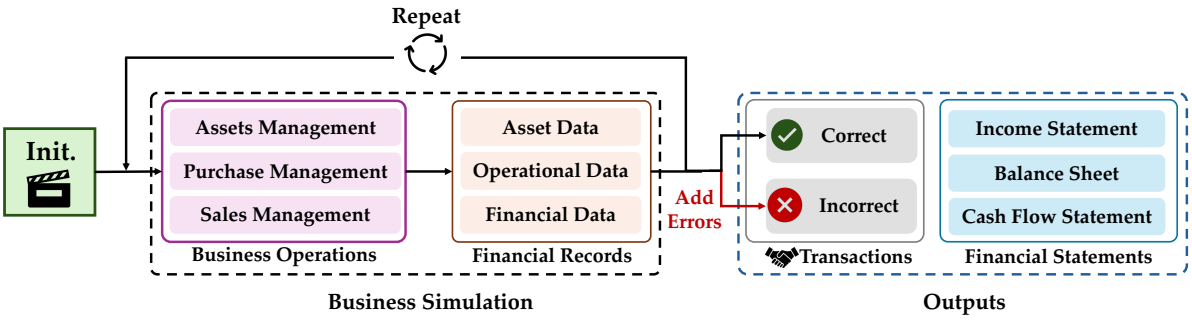
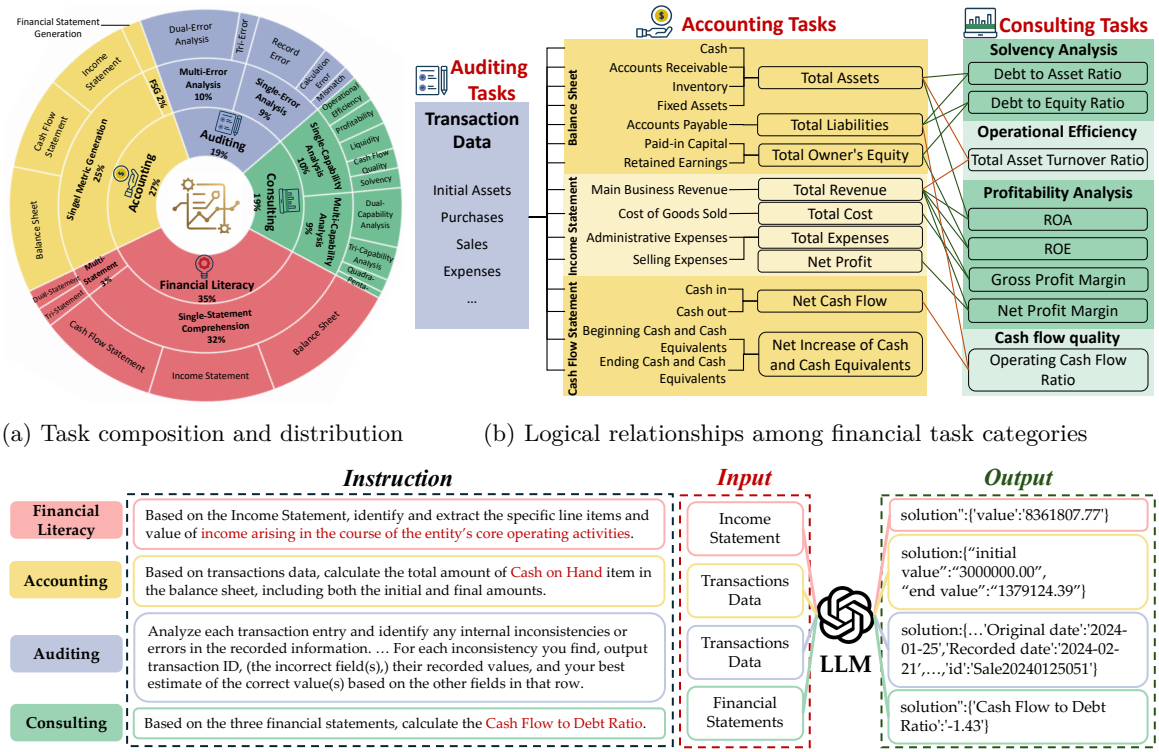


Figure 2: The workflow of *FinSim*.

Generation Process. Figure 2 illustrates the simulation workflow, where the architecture of *FinSim* follows a multi-stage process to generate transactions and financial statements.*FinSim* begins with initialization, where the simulator is configured to accurately model a specific type of company. Then the simulator proceeds to business simulation, where the business operations model financial activities, including assets management, purchase management, and sales management. These business operations further produce financial records, consisting of asset data, operational data, and financial data. This business simulation process will be repeated for the financial outputs. Transactions are derived from financial records, including both correct transactions and incorrect transactions, where the system adds typical errors to simulate real-world mistakes for auditing. For the financial statements, the income statement is generated by aggregating revenue and expense transactions from the financial data; the balance sheet combines asset positions from asset data with liability and equity information derived from operational data; and the cash flow statement synthesizes all cash-related transactions, categorizing them into operating, investing, and financing activities.

Design Principles and Validation. Despite generating synthetic data, *FinSim* ensures practical applicability through empirically grounded design and rigorous validation. The simulator implements a comprehensive accounting framework adhering to GAAP standards, incorporating approval workflows that mirror organizational hierarchies and asset valuation models following standard practices such as First-In-First-Out inventory costing and straight-line depreciation. Company archetypes and their operational parameters are calibrated from audited financial statements across industries. To capture real-world complexity, the simulator models market volatility through historical variance patterns, seasonal demand fluctuations via quarterly cycles, and industry-specific operational characteristics including supply chain dynamics and payment terms. Our validation methodology operates at multiple levels to ensure both accuracy and realism. Internal consistency checks verify double-entry bookkeeping integrity and mathematical accuracy across balance sheets, income statement, and cash flow statements. Compliance verification confirms adherence to revenue recognition principles, expense matching rules, and other GAAP requirements. External assessment by domain experts



(a) Task composition and distribution (b) Logical relationships among financial task categories

(c) Task structure examples across four financial domains

Figure 3: Task taxonomy and architecture.

with accounting and auditing backgrounds validates the authenticity of generated scenarios and transaction patterns. Finally, systematic error injection replicates common data imperfections, including data entry errors and timing discrepancies between transaction and settlement dates.

3.2 *FinSuite*: Financial Task Suite

FinSuite transforms the generated synthetic financial data into a comprehensive suite of 183 evaluable task instances across four domains: 64 in financial literacy, 49 in accounting, 35 in auditing, and 35 in consulting. Each task follows a unified structure comprising contextual inputs, domain-specific queries requiring multi-step reasoning, and verifiable outputs with traceable ground truth. The construction methodology is domain-adaptive: accounting tasks are built by selecting transaction subsets for statement generation, auditing tasks embed controlled error patterns for detection, consulting tasks aggregate multi-source statements for ratio analysis, and literacy tasks extract key figures for terminology assessment. The composition and proportions of tasks are shown in Figure 3a. The interdependencies among accounting, auditing, and consulting tasks are illustrated in Figure 3b, highlighting their interconnected nature which reflects real-world financial workflows. Figure 3c provides representative examples across task types to clarify the input-output formats.

Task Configuration. For each task, we construct an innovative three-dimensional metric system $\langle \alpha, \beta, \gamma \rangle$ to precisely characterize task features and complexity. Specifically: i) Computational base cardinality α: It refers to the number of fundamental data items required to solve the task, capturing the depth of the computational path; ii) Cross-source integration level β: It denotes the number of distinct input data sources involved, reflecting the complexity of data integration; iii) Output dimensionality breadth γ: It measures the number of target outputs, indicating the structural complexity of result generation. This multi-dimensional framework overcomes the limitations of traditional single-metric evaluations, enabling fine-grained error attribution. It also exhibits strong cross-domain adaptability, providing a unified complexity benchmark applicable to accounting, auditing, and consulting tasks, with potential for extension to other domains.

Financial Literacy Tasks. Financial literacy and statement comprehension are foundational for complex accounting tasks. To assess LLMs' competency in this domain, *FinSuite* introduces a Financial Literacy task using simulation-generated reports and definition-based queries requiring models to identify financial values without explicit terminology. Tasks are categorized by complexity based on input statements and output requirements, evaluating terminology understanding, cross-document reasoning, and data processing capabilities. This graduated framework serves as both a benchmark for foundational knowledge and a diagnostic tool for identifying knowledge gaps affecting downstream financial applications.

Accounting Tasks. Financial statement generation represents a core accounting function in contemporary financial reporting. *FinSuite* conceptualizes this fundamental process as the systematic transformation of granular transaction records into standardized financial statements through a comprehensive two-tiered framework, as illustrated in Figure 3a. The first tier involves generating precise disclosure items through both elementary computational operations focusing on discrete single transaction types and sophisticated complex computational operations requiring thorough cross-transaction analysis. The second tier methodically synthesizes these constituent components into comprehensive financial statements adhering to regulatory standards. This rigorous paradigm evaluates LLMs' technical capabilities in three critical dimensions: generating standardized values according to established accounting principles, achieving quantitative precision in multisource data integration, and maintaining methodological consistency in applying accounting standards. Through this structured approach, the framework establishes a robust benchmark for reliable financial outputs that effectively support subsequent auditing and consulting procedures in professional practice.

Auditing Tasks. Based on various areas of focus of audit in practice, *FinSuite* conceptualizes the audit task as a process of verifying transaction records within the financial audit framework, as shown in the Fugure 3b. The experimental paradigm constructs dual core components: generating realistic invoice-format transaction data to simulate authentic business environments and systematically embedding error samples within these records. Following error classifications in audits, twelve basic error types, grouped into three categories, were embedded into the dataset using randomized generation algorithms as primary evaluation targets. To deepen the measurement dimensions, error types were constructed in two levels: single error analysis and multi-error analysis, as shown in Figure 3a. This paradigm evaluates LLMs' performance in identifying audit errors of varying complexity and their semantic understanding of financial textual information.

Consulting Tasks. Analyzing a company's financial performance through quantitative financial indicators represents a critical service in professional financial consulting and investment decision-making. *FinSuite* conceptualizes the consulting task as a structured analytical framework based on established financial indicators, constructing a comprehensive diagnostic matrix comprising 18 essential indicators across five fundamental dimensions: profitability, operational efficiency, liquidity, solvency and cash flow quality. Each dimension captures distinct aspects of corporate financial health, ranging from profit generation capabilities to short-term liquidity management and long-term debt sustainability. By applying rigorous calculations on both individual and interrelated indicators, this paradigm methodically evaluates LLMs' technical accuracy in understanding financial indicator formulas, analytical traceability in data extraction from complex financial statements, and computational robustness in handling multi-step calculations, as illustrated in Figure 3c.

3.3 *FinEval*: LLM Evaluation

We introduce a unified evaluation framework with a designed prompt template to facilitate the evaluations.

Prompt Template. Our **FinMaster** prompt template adopts a standardized four-component structure designed for clarity and reproducibility. 1) Task description provides a concise statement that clearly defines the task name and its primary objective, establishing the foundational context for understanding the model. 2) Examples present detailed input-output pairs that demonstrate the expected solution format and reasoning approach through complete worked demonstrations. 3) The problem statement specifies the actual financial scenario to be solved, including all relevant data, market conditions, and contextual information necessary for a complete analysis. 4) Output instructions explicitly specify the structured JSON format requirements to ensure parsability across different model architectures. We employ a minimal instruction design philosophy: while we include a basic directive for step-by-step reasoning to ensure solution transparency, we deliberately avoid sophisticated prompt engineering techniques such as role-playing personas, task-specific optimization

hints, or elaborate multistage reasoning frameworks that might differentially benefit models with particular instruction-tuning characteristics. This design choice allows us to assess models' inherent financial reasoning capabilities rather than their responsiveness to complex prompting strategies.

```
finmaster_template = """
# <task_name> Task Description:
<task_description>

# Examples:
<in_context_examples>
# Problem to Solve:
{"problem": <task_to_solve>}

# Instruction:
Now please solve the above task. Reason step by step and present your answer in the "
    solution" field in the following json format:
```json
{"solution": "___" }
```
"""

example_and_solution = """{"problem": <example_problem>}
{"solution": <example_solution>}
"""
```

Completion with LLMs. We develop *FinEval*, a unified evaluation framework that operationalizes this template through three core components. 1) Prompt instantiation replaces template placeholders such as `<task_name>`, `<task_description>`, and `<task_to_solve>` with appropriate task-specific content drawn from our benchmark dataset, while preserving structural uniformity across all evaluations. 2) Unified execution provides consistent API interfaces across different LLM providers, equipped with robust error handling mechanisms, intelligent rate limiting to respect API constraints, and automatic retry logic to ensure reliable completion even under network instability or temporary service disruptions. 3) Response parsing extracts the solution field from diverse model outputs while correcting format issues such as missing quotation marks, handling incomplete responses through partial parsing strategies, and repairing JSON structures. This systematic pipeline eliminates technical inconsistencies and implementation artifacts, ensuring that observed performance differences genuinely reflect models' intrinsic financial reasoning capabilities and domain-specific knowledge rather than evaluation biases or technical implementation details.

4 Experimental Results

Evaluation Setup. We evaluate 7 representative LLMs spanning three major model families: GPT (OpenAI), Claude (Anthropic), and DeepSeek. Our comprehensive evaluation encompasses multiple dimensions: we validate the realism and logical consistency of **FinSim** through behavioral analysis, assess model performance across different task categories to understand their strengths and limitations in various financial reasoning scenarios, and conduct in-depth investigations into factors that influence financial reasoning capabilities. This multi-faceted approach enables both rigorous model comparison and systematic framework validation.

4.1 Validation of *FinSim* Simulation Process

To validate our *FinSim* framework's capability to accurately model real-world business operations, we conduct a manufacturing company simulation, whose results are presented in Figure 4. The simulation successfully reproduces the fundamental operational logic that drives real-world manufacturing business finances.

The **total assets** curve captures both predictable financial events, e.g., monthly operating expenses, payroll distributions, and quarterly tax payments, and stochastic elements, e.g., emergency equipment repairs and variable customer payment timing, that characterize actual business operations. The total asset growth exhibits a distinctive sawtooth pattern, in which cumulative sales gradually increase and are periodically offset by large expenditures, reflecting the financial rhythm observed in normally functioning enterprises. Notably, the simulation replicates capital investment behavior, where major fixed asset purchases appear as

substantial changes in asset composition while preserving overall value. The **weekly change** distribution further validates our approach. It accurately reproduces an asymmetric volatility profile, i.e., the positive accumulation periods (driven by sales and receivables) consistently outnumber negative outflow events (from expenses or unexpected costs), while negative events often exhibit larger magnitudes when they occur. This pattern directly aligns with real-world cash flow dynamics, where routine inflows sustain business operations but periodic outlays create occasional downward fluctuations.

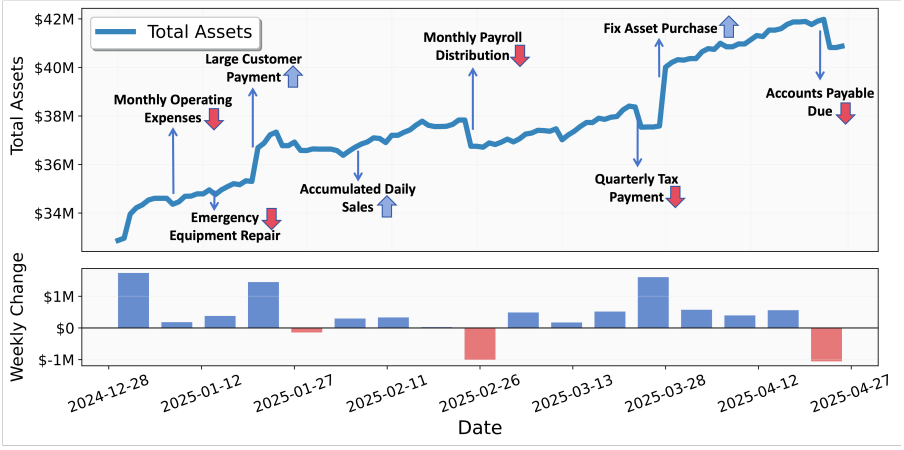


Figure 4: Simulation results for manufacturing company financial performance over time, showing total assets evolution.

This manufacturing case study demonstrates that our simulation framework effectively captures the complex interrelationships between revenue generation, operational requirements, and strategic investments that govern corporate financial trajectories. Hence, it provides compelling evidence that *FinSim* can reliably model authentic business operations within manufacturing sectors and possesses the adaptability to extend to other industries.

4.2 Analysis of Performance

The systematic evaluation of LLMs across financial tasks, presented in Table 3, reveals striking performance disparities that illuminate fundamental model limitations. All evaluations employ zero error tolerance without partial credits, reflecting the precision requirements of professional financial practice.

Financial Literacy. FinMaster first assesses the basic financial knowledge and statement understanding across different LLMs. As shown in Table 3, LLMs achieve strong performance in financial literacy tasks with a 96% average accuracy. Specifically, models including GPT-4.1, DeepSeek-V3, o3-mini, and Claude-3.7-Sonnet can reach nearly 100% accuracy, while GPT-4.1-nano and GPT-4o-mini show reduced performance, with an accuracy falling between 40% and 60%. This gap is especially due to their struggles with multistep reasoning and larger inputs. These results establish a crucial foundation for our comprehensive evaluation framework, demonstrating that advanced LLMs possess the requisite financial literacy to engage with more sophisticated analytical tasks. The significant performance disparity between models further underscores the value of this assessment in precisely calibrating expectations for different LLM capabilities across the financial domain.

Accounting. Accounting tasks expose critical weaknesses in multi-step financial reasoning. While all models achieve near-perfect accuracy (100%) on simple accounting tasks (tasks [1,1,1] through [3,3,3]), performance collapses dramatically as complexity increases. For instance, the performance on a moderate complex task [4,1,2] drops dramatically, i.e., 0% for all models tested in this study. This universal failure across all models indicates fundamental limitations in handling multi-step accounting workflows that involve complex calculations. We observe that the reasoning model o3-mini maintains a better average accuracy, i.e., 12.84%, with a much higher average token count, i.e., 10210.25, compared with other non-reasoning models, e.g., 10.86% average accuracy with 4,504 tokens for GPT-4.1. Most concerning, tasks requiring comprehensive statement generation ([14,1,1], [31,1,1], [37,1,2], [38,1,1]) yield 0-3.33% accuracy across nearly all models, demonstrating an inability to integrate multiple accounting principles simultaneously.

Auditing. Auditing tasks reveal pattern-dependent detection capabilities. o3-mini, DeepSeek-V3 and Claude-3.7-Sonnet dominate this category with average accuracies of 84.35% 67.33% and 68.01% respectively, substantially outperforming other models, e.g., GPT-4.1 (37.13%). However, granular analysis exposes an unexpected weakness: models perform better on multi-error scenarios than single-error detection. For example,

	Index	o3-mini		GPT-4o-mini		GPT-4.1-nano		GPT-4.1-mini		GPT-4.1		DeepSeek-V3		Claude-3.7	
		Acc(%)	Token	Acc(%)	Token	Acc(%)	Token	Acc(%)	Token	Acc(%)	Token	Acc(%)	Token	Acc(%)	Token
Financial Literacy	[1, 1, 1]	$100.00_{\pm 0.0}$	$246_{\pm 109}$	$96.48_{\pm 18.4}$	$246_{\pm 97}$	$93.22_{\pm 25.2}$	$194_{\pm 107}$	$97.96_{\pm 14.1}$	$146_{\pm 94}$	$100.00_{\pm 0.0}$	$20_{\pm 12}$	$100.00_{\pm 0.0}$	$229_{\pm 64}$	$100.00_{\pm 0.0}$	$160_{\pm 36}$
	[1, 3, 1]	$100.00_{\pm 0.0}$	$384_{\pm 195}$	$87.50_{\pm 33.1}$	$280_{\pm 68}$	$93.53_{\pm 24.6}$	$241_{\pm 70}$	$100.00_{\pm 0.0}$	$150_{\pm 69}$	$100.00_{\pm 0.0}$	$61_{\pm 56}$	$100.00_{\pm 0.0}$	$282_{\pm 76}$	$100.00_{\pm 0.0}$	$213_{\pm 42}$
	[2, 1, 2]	$100.00_{\pm 0.0}$	$377_{\pm 221}$	$89.67_{\pm 30.5}$	$288_{\pm 106}$	$88.42_{\pm 32.0}$	$245_{\pm 80}$	$95.83_{\pm 20.0}$	$163_{\pm 93}$	$99.67_{\pm 5.8}$	$63_{\pm 64}$	$97.00_{\pm 17.1}$	$383_{\pm 251}$	$99.00_{\pm 10.0}$	$201_{\pm 40}$
	[2, 3, 2]	$100.00_{\pm 0.0}$	$568_{\pm 191}$	$67.78_{\pm 47.0}$	$334_{\pm 74}$	$71.26_{\pm 45.5}$	$332_{\pm 121}$	$100.00_{\pm 0.0}$	$262_{\pm 84}$	$100.00_{\pm 0.0}$	$133_{\pm 85}$	$100.00_{\pm 0.0}$	$449_{\pm 286}$	$100.00_{\pm 0.0}$	$224_{\pm 32}$
	[3, 1, 3]	$100.00_{\pm 0.0}$	$471_{\pm 171}$	$80.00_{\pm 40.2}$	$336_{\pm 94}$	$64.96_{\pm 47.9}$	$360_{\pm 122}$	$100.00_{\pm 0.0}$	$256_{\pm 150}$	$100.00_{\pm 0.0}$	$105_{\pm 94}$	$100.00_{\pm 0.0}$	$845_{\pm 465}$	$100.00_{\pm 0.0}$	$243_{\pm 41}$
	[3, 3, 3]	$100.00_{\pm 0.0}$	$1143_{\pm 712}$	$91.11_{\pm 28.6}$	$364_{\pm 69}$	$40.23_{\pm 49.3}$	$401_{\pm 120}$	$100.00_{\pm 0.0}$	$378_{\pm 222}$	$100.00_{\pm 0.0}$	$170_{\pm 99}$	$100.00_{\pm 0.0}$	$574_{\pm 218}$	$100.00_{\pm 0.0}$	$257_{\pm 37}$
	[4, 1, 4]	$100.00_{\pm 0.0}$	$589_{\pm 266}$	$78.89_{\pm 41.0}$	$347_{\pm 79}$	$78.41_{\pm 41.4}$	$380_{\pm 93}$	$100.00_{\pm 0.0}$	$291_{\pm 53}$	$100.00_{\pm 0.0}$	$98_{\pm 71}$	$100.00_{\pm 0.0}$	$770_{\pm 253}$	$97.78_{\pm 14.8}$	$305_{\pm 63}$
	[5, 1, 5]	$100.00_{\pm 0.0}$	$648_{\pm 318}$	$91.11_{\pm 28.6}$	$353_{\pm 72}$	$87.50_{\pm 33.3}$	$410_{\pm 142}$	$100.00_{\pm 0.0}$	$124_{\pm 112}$	$100.00_{\pm 0.0}$	$88_{\pm 33}$	$100.00_{\pm 0.0}$	$1184_{\pm 351}$	$100.00_{\pm 0.0}$	$314_{\pm 41}$
	[6, 1, 6]	$100.00_{\pm 0.0}$	$775_{\pm 235}$	$60.00_{\pm 49.8}$	$422_{\pm 89}$	$68.97_{\pm 47.1}$	$447_{\pm 124}$	$100.00_{\pm 0.0}$	$144_{\pm 123}$	$100.00_{\pm 0.0}$	$93_{\pm 9}$	$100.00_{\pm 0.0}$	$1339_{\pm 430}$	$100.00_{\pm 0.0}$	$347_{\pm 27}$
	[7, 1, 7]	$100.00_{\pm 0.0}$	$628_{\pm 223}$	$100.00_{\pm 0.0}$	$503_{\pm 104}$	$100.00_{\pm 0.0}$	$544_{\pm 167}$	$100.00_{\pm 0.0}$	$119_{\pm 2}$	$100.00_{\pm 0.0}$	$120_{\pm 3}$	$100.00_{\pm 0.0}$	$730_{\pm 634}$	$100.00_{\pm 0.0}$	$416_{\pm 23}$
	Average	$100.00_{\pm 0.0}$	$620.33_{\pm 250.6}$	$82.90_{\pm 12.6}$	$358.56_{\pm 73.6}$	$77.03_{\pm 17.9}$	$373.33_{\pm 106.2}$	$99.54_{\pm 1.4}$	$209.67_{\pm 87.6}$	$99.96_{\pm 0.1}$	$103.44_{\pm 41.7}$	$99.67_{\pm 0.9}$	$728.44_{\pm 371.3}$	$99.64_{\pm 0.7}$	$280.00_{\pm 77.1}$
Accounting	[1, 1, 1]	$47.02_{\pm 50.0}$	$5673_{\pm 3167}$	$7.37_{\pm 26.1}$	$890_{\pm 500}$	$2.63_{\pm 16.0}$	$1146_{\pm 794}$	$24.91_{\pm 43.3}$	$1271_{\pm 901}$	$43.51_{\pm 49.6}$	$2044_{\pm 1730}$	$23.39_{\pm 42.4}$	$878_{\pm 560}$	$41.90_{\pm 49.4}$	$1015_{\pm 612}$
	[1, 1, 2]	$42.62_{\pm 49.5}$	$6798_{\pm 4524}$	$14.17_{\pm 34.9}$	$808_{\pm 401}$	$6.11_{\pm 24.0}$	$892_{\pm 694}$	$25.28_{\pm 43.5}$	$1608_{\pm 1259}$	$45.00_{\pm 49.8}$	$2694_{\pm 2426}$	$25.87_{\pm 43.9}$	$1123_{\pm 819}$	$30.45_{\pm 46.1}$	$1102_{\pm 648}$
	[2, 1, 1]	$55.00_{\pm 50.2}$	$6345_{\pm 5644}$	$26.67_{\pm 44.6}$	$984_{\pm 512}$	$36.36_{\pm 48.7}$	$1393_{\pm 942}$	$40.00_{\pm 49.4}$	$2555_{\pm 1879}$	$53.33_{\pm 50.3}$	$2556_{\pm 2406}$	$52.94_{\pm 50.4}$	$1221_{\pm 1327}$	$52.00_{\pm 50.5}$	$1018_{\pm 965}$
	[2, 1, 2]	$24.83_{\pm 43.4}$	$5943_{\pm 3507}$	$3.33_{\pm 18.0}$	$665_{\pm 206}$	$0.00_{\pm 0.0}$	$1057_{\pm 656}$	$22.00_{\pm 41.6}$	$1030_{\pm 699}$	$25.33_{\pm 43.6}$	$1872_{\pm 1888}$	$30.87_{\pm 46.4}$	$792_{\pm 603}$	$39.33_{\pm 49.0}$	$892_{\pm 428}$
	[4, 1, 1]	$28.33_{\pm 45.4}$	$10756_{\pm 3008}$	$0.00_{\pm 0.0}$	$1119_{\pm 444}$	$0.00_{\pm 0.0}$	$1392_{\pm 762}$	$1.67_{\pm 12.9}$	$2685_{\pm 1082}$	$6.67_{\pm 25.1}$	$3674_{\pm 1527}$	$1.85_{\pm 13.6}$	$1625_{\pm 685}$	$0.00_{\pm 0.0}$	$1595_{\pm 765}$
	[4, 1, 2]	$0.00_{\pm 0.0}$	$10533_{\pm 3384}$	$0.00_{\pm 0.0}$	$788_{\pm 153}$	$0.00_{\pm 0.0}$	$688_{\pm 608}$	$0.00_{\pm 0.0}$	$2219_{\pm 828}$	$0.00_{\pm 0.0}$	$5847_{\pm 1597}$	$0.00_{\pm 0.0}$	$1572_{\pm 913}$	$0.00_{\pm 0.0}$	$906_{\pm 307}$
	[7, 1, 1]	$0.00_{\pm 0.0}$	$11607_{\pm 2227}$	$0.00_{\pm 0.0}$	$1346_{\pm 441}$	$0.00_{\pm 0.0}$	$1412_{\pm 611}$	$0.00_{\pm 0.0}$	$3705_{\pm 1046}$	$0.00_{\pm 0.0}$	$6565_{\pm 1836}$	$0.00_{\pm 0.0}$	$1292_{\pm 897}$	$0.00_{\pm 0.0}$	$1310_{\pm 791}$
	[7, 1, 2]	$0.00_{\pm 0.0}$	$9914_{\pm 4595}$	$0.00_{\pm 0.0}$	$767_{\pm 157}$	$0.00_{\pm 0.0}$	$852_{\pm 702}$	$0.00_{\pm 0.0}$	$2172_{\pm 1030}$	$0.00_{\pm 0.0}$	$6678_{\pm 2167}$	$0.00_{\pm 0.0}$	$1805_{\pm 1203}$	$0.00_{\pm 0.0}$	$822_{\pm 244}$
	[8, 1, 1]	$0.00_{\pm 0.0}$	$11553_{\pm 2813}$	$0.00_{\pm 0.0}$	$1428_{\pm 507}$	$0.00_{\pm 0.0}$	$1547_{\pm 618}$	$0.00_{\pm 0.0}$	$3017_{\pm 1353}$	$0.00_{\pm 0.0}$	$6065_{\pm 1964}$	$0.00_{\pm 0.0}$	$1285_{\pm 927}$	$0.00_{\pm 0.0}$	$1220_{\pm 759}$
	[14, 1, 1]	$3.33_{\pm 18.3}$	$11896_{\pm 2461}$	$0.00_{\pm 0.0}$	$1357_{\pm 456}$	$0.00_{\pm 0.0}$	$2658_{\pm 1005}$	$0.00_{\pm 0.0}$	$3298_{\pm 973}$	$0.00_{\pm 0.0}$	$4525_{\pm 1225}$	$0.00_{\pm 0.0}$	$1586_{\pm 754}$	$0.00_{\pm 0.0}$	$1588_{\pm 669}$
	[31, 1, 1]	$0.00_{\pm 0.0}$	$13361_{\pm 2954}$	$0.00_{\pm 0.0}$	$1384_{\pm 557}$	$0.00_{\pm 0.0}$	$277_{\pm 472}$	$0.00_{\pm 0.0}$	$439_{\pm 439}$	$0.00_{\pm 0.0}$	$4525_{\pm 2388}$	$0.00_{\pm 0.0}$	$1214_{\pm 1181}$	$0.00_{\pm 0.0}$	$1483_{\pm 820}$
	[37, 1, 2]	$0.00_{\pm 0.0}$	$15881_{\pm 8846}$	$0.00_{\pm 0.0}$	$1537_{\pm 157}$	$0.00_{\pm 0.0}$	$506_{\pm 8}$	$0.00_{\pm 0.0}$	$681_{\pm 347}$	$0.00_{\pm 0.0}$	$4525_{\pm 2506}$	$0.00_{\pm 0.0}$	$2164_{\pm 1109}$	$0.00_{\pm 0.0}$	$1262_{\pm 550}$
	[38, 1, 1]	$0.00_{\pm 0.0}$	$7936_{\pm 4824}$	$0.00_{\pm 0.0}$	$1053_{\pm 312}$	$0.00_{\pm 0.0}$	$233_{\pm 20}$	$0.00_{\pm 0.0}$	$715_{\pm 501}$	$0.00_{\pm 0.0}$	$4525_{\pm 3831}$	$0.00_{\pm 0.0}$	$1237_{\pm 1125}$	$0.00_{\pm 0.0}$	$1213_{\pm 298}$
	Average	$12.84_{\pm 21.2}$	$10210.25_{\pm 3137.7}$	$3.68_{\pm 8.0}$	$1103.00_{\pm 295.3}$	$3.54_{\pm 10.0}$	$1075.58_{\pm 642.3}$	$7.41_{\pm 14.0}$	$2010.33_{\pm 1076.5}$	$10.86_{\pm 20.7}$	$4504.25_{\pm 1670.6}$	$9.29_{\pm 17.3}$	$1409.67_{\pm 374.9}$	$10.15_{\pm 20.2}$	$1200.92_{\pm 258.5}$
Auditing	[13, 1, 2]	$81.67_{\pm 39.0}$	$2670_{\pm 2079}$	$20.00_{\pm 40.3}$	$460_{\pm 214}$	$0.00_{\pm 0.0}$	$215_{\pm 29}$	$16.67_{\pm 37.6}$	$360_{\pm 232}$	$13.33_{\pm 34.3}$	$1262_{\pm 1135}$	$50.00_{\pm 50.4}$	$503_{\pm 985}$	$61.67_{\pm 49.0}$	$508_{\pm 115}$
	[13, 1, 3]	$93.00_{\pm 25.6}$	$1822_{\pm 1466}$	$24.00_{\pm 42.8}$	$456_{\pm 226}$	$0.00_{\pm 0.0}$	$196_{\pm 25}$	$29.67_{\pm 45.8}$	$465_{\pm 552}$	$45.67_{\pm 49.9}$	$849_{\pm 908}$	$72.33_{\pm 44.8}$	$310_{\pm 412}$	$76.33_{\pm 42.6}$	$542_{\pm 130}$
	[13, 1, 4]	$78.33_{\pm 41.4}$	$1858_{\pm 1161}$	$33.33_{\pm 47.3}$	$429_{\pm 178}$	$0.00_{\pm 0.0}$	$188_{\pm 14}$	$19.17_{\pm 39.5}$	$371_{\pm 330}$	$30.00_{\pm 46.0}$	$828_{\pm 857}$	$49.17_{\pm 50.2}$	$288_{\pm 153}$	$52.50_{\pm 50.1}$	$560_{\pm 112}$
	[13, 1, 5]	$87.62_{\pm 33.0}$	$1970_{\pm 1420}$	$36.67_{\pm 48.3}$	$478_{\pm 269}$	$0.95_{\pm 9.7}$	$197_{\pm 18}$	$38.10_{\pm 48.7}$	$402_{\pm 404}$	$53.81_{\pm 50.0}$	$698_{\pm 773}$	$87.62_{\pm 33.0}$	$300_{\pm 164}$	$89.05_{\pm 31.3}$	$549_{\pm 100}$
	[13, 1, 7]	$81.33_{\pm 39.1}$	$1948_{\pm 1240}$	$31.33_{\pm 46.5}$	$538_{\pm 212}$	$0.00_{\pm 0.0}$	$198_{\pm 10}$	$33.33_{\pm 47.3}$	$404_{\pm 455}$	$46.67_{\pm 50.1}$	$634_{\pm 463}$	$70.67_{\pm 45.7}$	$283_{\pm 127}$	$76.00_{\pm 42.9}$	$579_{\pm 95}$
	[13, 1, 9]	$84.17_{\pm 36.7}$	$2057_{\pm 1039}$	$21.67_{\pm 41.4}$	$543_{\pm 204}$	$0.00_{\pm 0.0}$	$212_{\pm 13}$	$25.83_{\pm 44.0}$	$371_{\pm 375}$	$33.33_{\pm 47.3}$	$766_{\pm 916}$	$74.17_{\pm 44.0}$	$294_{\pm 129}$	$52.50_{\pm 50.1}$	$613_{\pm 112}$
	[13, 1, 11]	$68.89_{\pm 46.5}$	$2527_{\pm 1396}$	$14.44_{\pm 35.4}$	$526_{\pm 215}$	$0.00_{\pm 0.0}$	$213_{\pm 8}$	$22.22_{\pm 41.8}$	$517_{\pm 716}$	$34.44_{\pm 47.8}$	$538_{\pm 237}$	$45.56_{\pm 50.1}$	$272_{\pm 67}$	$42.22_{\pm 49.7}$	$629_{\pm 72}$
	Average	$84.35_{\pm 7.6}$	$2054.17_{\pm 337.1}$	$27.83_{\pm 8.0}$	$484.00_{\pm 45.3}$	$0.16_{\pm 0.4}$	$201.00_{\pm 10.5}$	$27.13_{\pm 7.7}$	$395.50_{\pm 57.8}$	$37.13_{\pm 13.4}$	$839.50_{\pm 232.5}$	$67.33_{\pm 16.0}$	$329.67_{\pm 81.0}$	$68.01_{\pm 16.7}$	$558.50_{\pm 42.0}$
Consulting	[2, 1, 1]	$84.44_{\pm 36.3}$	$669_{\pm 263}$	$58.89_{\pm 49.3}$	$314_{\pm 81}$	$64.07_{\pm 48.1}$	$241_{\pm 89}$	$80.00_{\pm 40.1}$	$167_{\pm 40}$	$82.96_{\pm 37.7}$	$188_{\pm 50}$	$92.59_{\pm 26.2}$	$429_{\pm 328}$	$91.11_{\pm 28.5}$	$218_{\pm 42}$
	[2, 3, 1]	$95.00_{\pm 22.0}$	$814_{\pm 202}$	$85.00_{\pm 36.0}$	$349_{\pm 49}$	$72.41_{\pm 45.1}$	$346_{\pm 107}$	$66.67_{\pm 47.5}$	$234_{\pm 51}$	$95.00_{\pm 22.0}$	$219_{\pm 56}$	$96.67_{\pm 18.1}$	$335_{\pm 314}$	$98.33_{\pm 12.9}$	$250_{\pm 35}$
	[3, 3, 1]	$86.11_{\pm 34.7}$	$839_{\pm 189}$	$57.22_{\pm 49.6}$	$383_{\pm 61}$	$64.04_{\pm 48.1}$	$348_{\pm 91}$	$91.67_{\pm 27.7}$	$270_{\pm 29}$	$75.00_{\pm 43.4}$	$366_{\pm 113}$	$96.11_{\pm 19.4}$	$350_{\pm 84}$	$95.00_{\pm 21.9}$	$291_{\pm 27}$
	[4, 3, 2]	$50.00_{\pm 50.3}$	$1178_{\pm 420}$	$21.11_{\pm 41.0}$	$465_{\pm 64}$	$24.14_{\pm 43.0}$	$366_{\pm 85}$	$41.11_{\pm 49.5}$	$312_{\pm 64}$	$37.78_{\pm 48.8}$	$327_{\pm 92}$	$81.11_{\pm 39.4}$	$998_{\pm 272}$	$83.33_{\pm 37.5}$	$317_{\pm 28}$
	[5, 3, 2]	$86.67_{\pm 34.3}$	$1230_{\pm 230}$	$36.67_{\pm 48.6}$	$472_{\pm 64}$	$44.83_{\pm 50.2}$	$394_{\pm 74}$	$85.00_{\pm 36.0}$	$340_{\pm 61}$	$46.67_{\pm 50.3}$	$505_{\pm 162}$	$83.33_{\pm 37.6}$	$1404_{\pm 539}$	$98.33_{\pm 12.9}$	$356_{\pm 31}$
	[6, 1, 1]	$86.67_{\pm 34.6}$	$1056_{\pm 296}$	$40.00_{\pm 49.8}$	$434_{\pm 60}$	$63.33_{\pm 49.0}$	$329_{\pm 65}$	$93.33_{\pm 25.4}$	$293_{\pm 52}$	$90.00_{\pm 30.5}$	$325_{\pm 63}$	$100.00_{\pm 0.0}$	$535_{\pm 320}$	$100.00_{\pm 0.0}$	$263_{\pm 31}$
	[6, 3, 3]	$36.67_{\pm 48.6}$	$1291_{\pm 212}$	$15.00_{\pm 36.0}$	$598_{\pm 60}$	$17.24_{\pm 38.1}$	$463_{\pm 79}$	$63.33_{\pm 48.6}$	$458_{\pm 63}$	$46.67_{\pm 50.3}$	$424_{\pm 117}$	$91.67_{\pm 27.9}$	$1080_{\pm 221}$	$86.67_{\pm 34.3}$	$420_{\pm 28}$
	[7, 3, 3]	$23.33_{\pm 43.0}$	$1425_{\pm 295}$	$0.00_{\pm 0.0}$	$570_{\pm 58}$	$3.45_{\pm 18.6}$	$452_{\pm 100}$	$23.33_{\pm 43.0}$	$505_{\pm 77}$	$10.00_{\pm 30.5}$	$435_{\pm 84}$	$23.33_{\pm 43.0}$	$1114_{\pm 220}$	$10.00_{\pm 30.5}$	$393_{\pm 43}$
	[8, 3, 2]	$53.33_{\pm 50.7}$	$1187_{\pm 286}$	$6.67_{\pm 25.4}$	$528_{\pm 84}$	$10.34_{\pm 31.0}$	$458_{\pm 97}$	$70.00_{\pm 46.6}$	$444_{\pm 47}$	$43.33_{\pm 50.4}$	$514_{\pm 158}$	$73.33_{\pm 45.0}$	$1624_{\pm 518}$	$90.00_{\pm 30.5}$	$343_{\pm 23}$
	[10, 3, 4]	$0.00_{\pm 0.0}$	$1568_{\pm 241}$	$0.00_{\pm 0.0}$	$714_{\pm 71}$	$3.45_{\pm 18.6}$	$641_{\pm 99}$	$23.33_{\pm 43.0}$	$617_{\pm 78}$	$6.67_{\pm 25.4}$	$755_{\pm 201}$	$16.67_{\pm 37.9}$	$1023_{\pm 296}$	$0.00_{\pm 0.0}$	$546_{\pm 45}$
	[11, 3, 3]	$53.33_{\pm 50.7}$	$2064_{\pm 483}$	$16.67_{\pm 37.9}$	$702_{\pm 86}$	$6.90_{\pm 25.8}$	$651_{\pm 180}$	$56.67_{\pm 50.4}$	$602_{\pm 69}$	$23.33_{\pm 43.0}$	$724_{\pm 302}$	$43.33_{\pm 50.4}$	$1900_{\pm 620}$	$76.67_{\pm 43.0}$	$473_{\pm 43}$
	[12, 3, 5]	$80.00_{\pm 40.7}$	$1923_{\pm 190}$	$23.33_{\pm 43.0}$	$849_{\pm 91}$	$24.14_{\pm 43.5}$	$839_{\pm 263}$	$46.67_{\pm 50.7}$	$869_{\pm 95}$	$36.67_{\pm 49.0}$	$811_{\pm 171}$	$80.00_{\pm 40.7}$	$1307_{\pm 374}$	$66.67_{\pm 48.0}$	$628_{\pm 48}$
	[14, 3, 5]	$76.67_{\pm 43.0}$	$1713_{\pm 351}$	$13.33_{\pm 34.6}$	$762_{\pm 73}$	$31.03_{\pm 47.1}$	$696_{\pm 123}$	$76.67_{\pm 43.0}$	$735_{\pm 99}$	$56.67_{\pm 50.4}$	$592_{\pm 98}$	$73.33_{\pm 45.0}$	$1088_{\pm 156}$	$93.33_{\pm 25.4}$	$498_{\pm 92}$
	[14, 3, 6]	$13.33_{\pm 34.6}$	$2176_{\pm 427}$	$0.00_{\pm 0.0}$	$874_{\pm 66}$	$3.45_{\pm 18.6}$	$708_{\pm 92}$	$3.33_{\pm 18.3}$	$807_{\pm 114}$	$0.00_{\pm 0.0}$	$647_{\pm 99}$	$66.67_{\pm 48.0}$	$1194_{\pm 148}$	$36.67_{\pm 49.0}$	$592_{\pm 102}$
	[15, 3, 7]	$0.00_{\pm 0.0}$	$2132_{\pm 358}$	$0.00_{\pm 0.0}$	$898_{\pm 99}$	$0.00_{\pm 0.0}$	$778_{\pm 180}$	$0.00_{\pm 0.0}$	$918_{\pm 133}$	$0.00_{\pm 0.0}$	$574_{\pm 77}$	$0.00_{\pm 0.0}$	$1101_{\pm 200}$	$0.00_{\pm 0.0}$	$580_{\pm 63}$
	[17, 3, 6]	$66.67_{\pm 48.0}$	$2406_{\pm 494}$	$0.00_{\pm 0.0}$	$909_{\pm 78}$	$13.79_{\pm 35.1}$	$903_{\pm 155}$	$63.33_{\pm 49.0}$	$923_{\pm 123}$	$13.33_{\pm 34.6}$	$812_{\pm 242}$	$70.00_{\pm 46.6}$	$1482_{\pm 417}$	$83.33_{\pm 37.9}$	$615_{\pm 94}$
	[21, 3, 7]	$10.00_{\pm 30.5}$	$2722_{\pm 607}$	$0.00_{\pm 0.0}$	$999_{\pm 64}$	$3.45_{\pm 18.6}$	$965_{\pm 147}$	$23.33_{\pm 43.0}$	$1027_{\pm 230}$	$6.67_{\pm 25.4}$	$916_{\pm 287}$	$56.67_{\pm 50.4}$	$1438_{\pm 393}$	$46.67_{\pm 50.7}$	$784_{\pm 66}$
	Average	$51.11_{\pm 33.3}$	$1607.75_{\pm 601.5}$	$19.69_{\pm 25.5}$	$656.62_{\pm 219.1}$	$24.12_{\pm 25.5}$	$583.56_{\pm 224.3}$	$51.74_{\pm 29.8}$	$584.62_{\pm 276.5}$	$36.74_{\pm 31.9}$	$559.12_{\pm 217.8}$	$65.76_{\pm 30.0}$	$1123.31_{\pm 450.5}$	$66.56_{\pm 35.5}$	$459.31_{\pm 161.1}$

Table 3: Full results on *FinSuite* benchmark across different models. The token numbers represent the completion tokens generated by each model. Results are presented with accuracy percentages (Acc%) and standard deviations across four main evaluation domains: Financial Literacy, Accounting, Auditing, and Consulting. Average performance metrics are calculated for each domain to facilitate model comparison.

on task [13,1,5] (5 simultaneous errors), Claude-3.7-Sonnet achieves 89.05% accuracy, while on task [13,1,1] (single error), accuracy drops to 68.89%. This pattern is consistent across models, e.g., DeepSeek-V3 scores 49.17% on [13,1,4] (4 errors) versus 45.56% on [13,1,1] (1 error). The token efficiency also varies dramatically: o3-mini uses 2,054 tokens on average for auditing tasks compared to GPT-4.1-nano's 201 tokens, yet achieves superior accuracy (84.35% vs 0.16%), suggesting a trade-off between the tokens and performance.

Consulting. Consulting tasks demonstrate inconsistent cross-statement analytical capabilities. Performance variability is extreme within this category. On basic consulting tasks like [2,1,1], models perform reasonably well (o3-mini: 84.44%, Claude-3.7-Sonnet: 91.11%), but complex multi-statement analysis in tasks like [7,3,3] causes severe degradation (o3-mini: 23.3%, GPT-4o-mini: 0%, GPT-4.1-nano: 3.45%, Claude-3.7-Sonnet: 10%). Noteworthy, GPT-4.1-nano achieves 24.12% average accuracy in consulting despite catastrophic failure in accounting (3.54%). Specific tasks reveal this specialization: on [10,3,4], GPT-4.1-nano scores 3.45% while GPT-4.1 achieves only 6.67%, yet on [2,1,1], the relationship reverses (64.07% vs 82.96%). DeepSeek-V3 demonstrates the most consistent high-level performance with 65.76% average accuracy, though it consumes 1,123 tokens on average with nearly double GPT-4.1's 559 tokens, a trade-off between tokens and performance.

Token Efficiency. Token efficiency does not necessarily correlate with accuracy. For simple tasks, nonreasoning models achieve comparable performance with substantially fewer tokens: GPT-4.1 reaches 99. 96% accuracy on financial literacy tasks using 103 tokens versus o3-mini's 620 tokens for 100% accuracy, demonstrating that reasoning models offer no advantage on straightforward calculations. For moderately difficult tasks, DeepSeek-V3 uses 329 tokens to achieve 74.17% auditing accuracy, outperforming GPT-4.1-nano's 201 tokens for only 27.13% accuracy, representing a 2.7x performance gain for 1.6x token cost. However, for extremely difficult tasks, the performance advantage diminishes despite exponentially higher token consumption: o3-mini explodes to 10,210 tokens for only 12.84% accounting accuracy compared to GPT-4.1 10. 86% at 4,504 tokens. This pattern suggests that simple tasks favor lightweight models for efficiency, moderate tasks justify reasoning models' computational overhead, but beyond a complexity threshold, additional tokens cannot compensate for architectural limitations in multi-step financial reasoning.

Takeaways

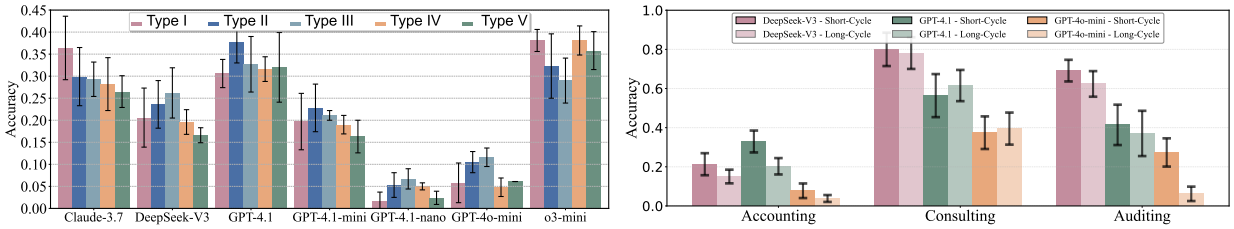
- **FinMaster** effectively reveals LLMs' limitations in real-world financial services.
- LLMs achieves good performance on the financial literacy (99%-100%), simple auditing and consulting tasks ($> 80\%$).
- LLMs fails to conduct simple accounting tasks, even for o3-mini and Claude-3.7-Sonnet, and show the gaps for complex auditing and consulting tasks.
- o3-mini performs best in the accounting and auditing tasks, while DeepSeek-V3 and Claude-3.7-Sonnet performs best in the consulting tasks, suggesting the fundamental differences of these models.
- Models with more tokens do not necessarily translate to better performance, depending on the difficulties of tasks.

4.3 Ablations

Different Companies Comparison. *FinSim* designs five company types with unique operational characteristics to explore how organizational settings affect model performance. As shown in Figure 5a, high-accuracy models (Claude-3.7-Sonnet, DeepSeek-V3) show consistent performance with shorter error bars, while lower-accuracy models (GPT-4.1-mini, GPT-4.1-nano) exhibit greater variability with longer error bars. Interestingly, this pattern persists across all company archetypes despite their different financial structures. The results reveal that business operations impact model effectiveness and stability, with strong models maintaining stable performance across all company types while weaker models are more sensitive to organizational differences. These findings highlight the need to match model choice with real-world business complexity.

Companies Operation Duration Comparison. In the simulation process, *FinSim* uses transaction volume as a proxy for operational time periods, comparing short-cycle scenarios with 200 transactions versus long-cycle scenarios with 400 transactions per company. As shown in Figure 5b, consulting tasks

demonstrate remarkable stability: GPT-4o-mini maintains 37%-39% performance across both cycles, and GPT-4.1 improves slightly from 56% to 61% in long cycles. In contrast, transaction processing tasks exhibit notable performance degradation: DeepSeek-V3's accuracy in accounting decreases from 21% to 15%, while its auditing performance declines from 69% to 62% in long cycles. These findings indicate that although models maintain consistent performance in analyzing standardized financial statements, their ability to perform reasoning and calculation declines as operational periods expand.



(a) Performance over companies.

(b) Model accuracy for different operation time.

Figure 5: Ablation study for the comparisons of different companies and companies operation duration.

Takeaways

- **FinMaster** designed tasks capture universal business logic applicable across various industry contexts.
- Strong models maintain consistency across company types while weak models show high variability.
- Increasing input transactions weaken LLMs' reasoning and multi-step calculation ability.
- Model output variance in LLMs stays consistent across cycles, suggesting inherent uncertainty.

4.4 Analysis of Reasoning Failure Cases

We analyze the errors made by LLMs across all tasks to identify their core weaknesses, with insights intended to inform future domain-specific optimization and model improvement. Based on a systematic error review, four major types of failure reasons for LLMs are summarized. As displayed in Table 4, all examples are drawn from o3-mini, which show relatively transparent and interpretable reasoning processes.

The failure reasons include: i) **Domain Knowledge Gap.** Although LLMs generally achieve high accuracy in financial literacy tasks, they still show significant misunderstandings of professional financial concepts in real applications. As shown in Table 4, the model misclassified interest receivable as operating revenue and miscalculated ROA by using end-period rather than average assets, revealing gaps in applying basic accounting and financial principles. This also indicates a disconnection between terminology recognition and practical knowledge application;ii) **Critical Data Missing.** LLMs tend to omit essential financial data or adjustment items during calculation. For the example displayed in Table 4, the model failed to include the "bank-to-cash transfer" when computing total cash outflow and ignored non-cash adjustments in calculating operating cash flow. These errors reflect limited awareness of data completeness and weak ability to track dynamic data flows and cross-check accounting relationships in multi-step reasoning;iii) **Floating Point Error.** When required to retain two decimal places, LLMs often truncate rather than correctly round the number. Such numerical inaccuracies are critical in financial domain, where even minor deviations may cause reconciliation failures or invalid audit results;iv) **Reasoning Consistency Error.** LLMs frequently produce logically inconsistent steps in multi-stage calculations. For example, intermediate figures or formulas used in later steps sometimes conflict with the model's own earlier outputs. This issue indicates a temporal break in the reasoning chain and reveals the lack of self-monitoring and verification mechanisms during generation. In summary, LLMs exhibit both conceptual weaknesses, e.g., incomplete understanding of key principles, and procedural errors, e.g., logical inconsistency and data missing, along with computational instability in numerical precision. These errors undermine the model's usability in financial decision-making and audit support, highlighting structural deficiencies in logical consistency, data integrity, and numerical accuracy.

Failure Reason 1: Domain Knowledge Gap

Case (1): ... All Sale transactions were summed to yield a Main Business Revenue of 6936375.28. Adding positive interest-earnin "Interest Receivables" (net 1735.58) gives Total Revenue of 6938110.86. The cost side was computed by summing the "cost_amount" on all sale rows (totaling 5,700,000.00) which is taken as the Cost of Goods Sold (and Total Cost)...
// **Domain knowledge gap in accounting ontology**: Treats internal transfers (e.g., Cash to Bank) as pure additions without recognizing they are neutral to total liquidity, leading to double-counting in bank deposits and misrepresenting cash equivalents. This exemplifies domain knowledge gap by incorrectly including Interest Receivables (a non-operating accrual asset) as part of Main Business Revenue, distorting the revenue ontology and leading to inflated Total Revenue, as Interest should be a separate non-operating income item in financial statements.

Case (2): ... Net Profit=1,095,371.35, ..., Total Assets=14,823,650.85,... Calculate the Return on Assets (ROA) using the formula: ROA = (Net Profit / Total Assets)×100%: ROA= $(1,095,371.35/14,823,650.85) \times 100\% \approx 7.39\%$. Thus, the ROA is approximately 7.39%...
// **Domain knowledge gap in financial ratio** : The correct formula should be ROA = NET PROFIT / AVERAGE TOTAL ASSET, i.e., the denominator should be the sum of total assets at the beginning and end of the period divided by two. However, the denominator in this case is incorrectly using only the total assets at the end of the period.

Failure Reason 2: Critical Data Missing

Case (1): ... The sum of all cash outflows (all payments made "by Cash"—for expenses, cash purchases, and all "Cash to bank transfer" items) is about 6,361,868.78 and all "Cash to bank transfer" items) is about 6,361,868.78. Subtracting, the ending balance is 9,333,895.58 – 6,361,868.78 = 2,972,026.80. Thus, in our simulation the cash on hand starts at 3,000,000.00...
// **Key evidence missing**: Omits bidirectional transfers in cash simulation, eliding key ledger inputs that neutralize net cash position, resulting in overstated ending Cash on Hand. The analysis fails to properly account for inter-account movements that should be treated as internal transfers rather than true cash flows, creating artificial inflation in liquidity metrics.

Case (2): ... Operating outflows (expenses) – the sum of all paid operating expense amounts comes out to about 3,285,464.64. Operating inflows (sales receipts) – the sum of all paid sales amounts comes out to about 7,514,065.92. Subtracting expenses from receipts we get a net operating cash flow of 7,514,065.92 − 3,285,464.64 = 4,228,601.28.
// **Omit adjustment information**: Omits changes in working capital (e.g., AR increases, inventory purchases) and non-cash adjustments (e.g., depreciation, amortization) in direct method cash flow. The calculation fails to incorporate key reconciling items such as changes in accounts payable, resulting in an oversimplified representation of operational cash generation capacity.

Failure Reason 3: Floating Point Error

Case (1): ...To calculate ROA, we first compute the average total assets and then divide the net profit by this average... Calculate the Return on Assets (ROA): ROA = (Net Profit/Average Total Assets)×100 = (–1,342,040.76/14,043,024.825)×100 ≈ –9.55% . Thus,the ROA is approximately –9.55%.The solution in the required format is : {"solution": {"Return on Assets (ROA)": "-9.55%"}...
// **Floating point error**: Truncation in average assets (14,043,024.825 vs. precise 14,043,024.825) and inexpert rounding yield -9.55% instead of correct -9.56%, due to decimal precision loss.

Case (2): ... Step 2: Calculate the average Owner's Equity. Average Owner's Equity = (Beginning Owner's Equity + Ending Owner's Equity) / 2 = (13,000,000.00 + 11,657,959.24) / 2 = 24,657,959.24 / 2 = 12,328,979.62. Step 3: Calculate ROE. ROE = (Net Profit / Average Owner's Equity) × 100 = (−1,342,040.76 / 12,328,979.62) × 100 ≈ −10.88%.
// **Rounding heuristic error**: Average equity 12,328,979.62 leads to ≈ −10.88% but correct is −10.89% due to floating point deviation in $(-1,342,040.76/12,328,979.62) \times 100$, subtle but undermining ratio fidelity. This precision inconsistency, while seemingly minor, propagates through financial analysis and could significantly impact decision-making.

Failure Reason 4: Reasoning Consistency Error

Case (1): ... Total Sale − profit + interest income ≈ 1,721,037 + 4,535 = 1,725,572. Total Expenses ≈ 2,227,153. Then the net income (profit or loss) in the period is Net income ≈ 1,725,572 − 2,227,153 ≈ −501,580. If we assume that no previous period's retained earnings remain so that the beginning...
// **Logic chain incoherence**: Sums sales profits to 1,721,037 but later implies different total in RE calculation; net income -501,580 inconsistent with final answer implying different loss, eroding chain robustness. Fails to reconcile intermediate calculations with final figures, demonstrating poor numerical tracking and verification throughout the solution process.

Case (2): ... So that finally Total Assets End Value = initial (13,000,000) + net change (≈ 10,268,950) ≈ 23,268,950. Because of the many small transactions the numbers require keeping track with many "steps." With an exact careful summing the final answer (to two−decimal accuracy) is: Total Assets Initial Value = 13,000,000.00. Total Assets End Value = 23,268,949.86...
// **Internal incoherence**: Initial net change ≈ 10,268,950 leads to ≈ 23,268,950, but final precise value 23,268,949.86 without reconciling the 0.14 discrepancy. This inconsistency between rounded intermediate calculations and final precise values creates ambiguity about the computational pathway.

Table 4: Examples of Failure Cases for o3-mini.

5 Conclusion

FinMaster is a comprehensive benchmark for evaluating LLMs on real-world financial management across accounting, auditing, and advisory domains. The benchmark comprises three modules: *FinSim* synthesizes realistic financial data, *FinSuite* encompasses 183 diverse tasks, and *FinEval* provides a unified assessment framework. Evaluation of models such as Claude-3.7-Sonnet, DeepSeek-V3, and o3-mini uncovers a critical gap: while achieving 96% accuracy on foundational tasks, performance deteriorates to 40% on complex scenarios demanding multi-source integration and domain-specific reasoning. Computational error propagation in multi-step reasoning constitutes the primary bottleneck. **FinMaster** provides the first benchmark designed to guide LLM development toward reliable financial management capabilities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Edward I Altman, Małgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. *Journal of international financial management & accounting*, 28(2):131–171, 2017.

Elaine Biech. *The new business of consulting: the basics and beyond*. John Wiley & Sons, 2019.

Miriam Bruhn, Dean Karlan, and Antoinette Schoar. The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico. *Journal of Political Economy*, 126(2):635–687, 2018.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

Barry J Epstein, Ralph Nach, and Steven M Bragg. *Wiley GAAP 2010: Interpretation and application of generally accepted accounting principles*. John Wiley & Sons, 2009.

Jayne Godfrey, Allan Hodgson, Ann Tarca, Jane Hamilton, and Scott Holmen. *Accounting*. John Wiley & Sons, Inc, 2010.

Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu-tao Xia, Dawei Cheng, and Changjun Jiang. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*, 2025.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Sathvik Joel, Jie Wu, and Fatemeh Fard. A survey on llm-based code generation for low-resource and domain-specific programming languages. *ACM Transactions on Software Engineering and Methodology*, 2024.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*, 2023.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*, 2023.

Debbi Chyntia Ovami and Iskandar Muda. Data analytics and its implication on auditing. In *12th International Conference on Green Technology (ICGT 2022)*, pp. 93–101. Atlantis Press, 2023.

ReportLinker. Financial services global market report 2023, April 2023. URL `https://www.reportlinker.com/p06277918/?utm_source=GNW`.

Luis Paulo Guimarães dos Santos, Anderson José Freitas de Cerqueira, and César Valentim de Oliveira Carvalho. An experimental analysis of the effect of recordkeeping over direct reciprocity. *Revista Contabilidade & Finanças*, 32(86):359–375, 2020.

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can llms master math? investigating large language models on math stack exchange. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 2316–2320, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

EARNING PER SHARE. Income statement. *Group*, 1:01–07, 1995.

Mark T Soliman. The use of dupont analysis by market participants. *The accounting review*, 83(3):823–853, 2008.

Carola Westermeier. Money is data–the platformization of financial transactions. *Information, Communication & Society*, 23(14):2047–2063, 2020.

Gerald I White, Ashwinpaul C Sondhi, and Dov Fried. *The analysis and use of financial statements*. John Wiley & Sons, 2002.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

Noga Ben Yoash, Meni Brief, Oded Ovadia, Gil Shenderovitz, Moshik Mishaeli, Rachel Lemberg, and Eitam Sheetrit. SECQUE: A benchmark for evaluating real-world financial analysis capabilities. *arXiv preprint arXiv:2504.04596*, 2025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. 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, 2024.

Hanyu Zhang, Boyu Qiu, Yuhao Feng, Shuqi Li, Qian Ma, Xiyuan Zhang, Qiang Ju, Dong Yan, and Jian Xie. Baichuan4-finance technical report. *arXiv preprint arXiv:2412.15270*, 2024a.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. Finagent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv e-prints*, pp. arXiv–2402, 2024b.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. FinanceMATH: Knowledge-intensive math reasoning in finance domains. In *ACL*, pp. 12841–12858, 2024a.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. DOCMATH-EVAL: Evaluating math reasoning capabilities of llms in understanding financial documents. In *ACL*, pp. 16103–16120, 2024b.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *ACL*, pp. 3277–3287, 2021.

Jie Zhu, Qian Chen, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. DianJin-R1: Evaluating and enhancing financial reasoning in large language models. *arXiv preprint arXiv:2504.15716*, 2025.

A Frequently Asked Questions (FAQs)

A.1 Why Using Simulated Data is Enough?

Simulated data addresses several critical challenges in financial LLM research:

- Privacy and Compliance: Real financial data often contains sensitive information, e.g., client records and proprietary transactions, subject to various strict regulations. Simulated transaction records and financial statements generated by *FinSim* can avoid exposing confidential information to eliminate privacy risks while replicating real-world complexity.
- Scalability and Diversity: Financial data is inherently sensitive, regulated, and fragmented across institutions, making real-world datasets difficult to obtain and limited in scope. *FinSim* can dynamically generate limitless datasets for diverse types of companies and market conditions, enabling robust LLM evaluation without data scarcity.
- Evaluation Control: Simulated data makes evaluation controllable. i) Ground truth control: *FinSim* establishes precise, verifiable ground truth for each specific task, which is critical for evaluating the accuracy and robustness of LLMs in financial tasks. ii) Complexity control: *FinSim* can systematically modulate task difficulty, which allows benchmarking LLMs' scalability across operational environments. iii) Error injection control: *FinSim* allows precise manipulation of variables, e.g., for auditing tasks, *FinSim* can simulate transactions with injecting errors, especially for the errors that are impractical to replicate with real data, to evaluate LLMs' anomaly detection capabilities without exposing real-world misconduct.
- Generalizability: Financial tasks, e.g., accounting, auditing and consulting, rely on standardized formats rather than unpredictable market dynamics, e.g., trading, allowing *FinSim* to replicate realistic data structures, logical relationships, and error patterns that mirror real-world complexity. This ensures LLM performance on simulated data transfers reliably to real-world scenarios, with minimal out-of-distribution (OOD) divergence.

A.2 Why Focusing on Accounting, Auditing and Consulting?

We focus on accounting, auditing and consulting due to three fundamental considerations that collectively establish them as both critical and uniquely positioned for LLM-driven transformation in financial services:

- Critical roles in financial workflows: Accounting, auditing and consulting underpin global financial systems, where accounting ensures accurate record-keeping and compliance, auditing provides the verification mechanism to ensure the reliability and integrity of financial reporting, and consulting translates financial data into actionable business insights and investment decisions. Their temporal dependencies create a comprehensive evaluation framework where LLMs should demonstrate proficiency across data processing, verification, and strategic interpretation.
- High automation potential: Accounting and auditing involve rule-based but labor-intensive processes, e.g., intercompany reconciliations and anomaly detection in ledgers, ideal for LLM automation to reduce human effort and errors. Consulting leverages LLMs for rapid insights from financial statements or regulatory documents.
- Under-explored complexity in existing LLM benchmarks: Existing benchmarks, e.g., FinQA (Chen et al., 2021), prioritize narrow tasks like numerical question answering from static reports, which only require single-step reasoning and fail to capture the full spectrum of financial reasoning, i.e., from data processing to strategic decision-making. We present **FinMaster**, which involves tasks requiring multi-step reasoning across the entire financial workflow, to fill this under-explored complexity gaps in existing LLM benchmarks.

A.3 Model Selection

Due to budget constraints, our evaluation focuses on a curated set of representative models. Specifically, we select online advanced non-reasoning models, e.g., GPT-4o-mini, Claude-3.7-Sonnet, and DeepSeek-V3, and online reasoning model, i.e., o3-mini. For more recent models, we plan to incorporate them in the next update of our benchmark.

A.4 Discussion about Limitations and Future Works

Multimodal Financial Analysis. Current financial tasks in **FinMaster** are text-based, but real-world financial analysis sometimes involves multimodal data, e.g., charts and scanned documents. Therefore, expanding *FinSim* to generate multimodal financial data and considering multimodal LLMs (MLLMs) can extend **FinMaster** to include multimodal financial reasoning, where a multimodal version of **FinMaster** would better simulate real-world scenarios.

Retrieval Augmented Generation (RAG) for Financial Reasoning. Financial reasoning often requires access to large-scale datasets, while current LLMs struggle with limited context window size and long-context retention. Therefore, exploring dynamic data retrieval, i.e., integrating RAG to fetch relevant financial data, provides a promising direction for future works to reduce hallucination and improve accuracy.

Domain-Specific Training for Financial Expertise. General-purpose LLMs lack deep financial knowledge, leading to misinterpretation of financial concepts. **FinMaster** provides high-quality financial training data through *FinSim*, supporting LLM specialized fine-tuning, which can improve LLM financial reasoning capability and even develop finance domain-specific foundation models.

A.5 Negative Impacts

We do not foresee any negative impacts.

B Related Work

Financial Benchmarks. We provide a review of existing financial benchmarks for LLMs evaluation. FinQA (Chen et al., 2021) introduces a novel dataset for complex financial QA, requiring models to interpret hybrid data (text/tables) from financial reports, perform multi-step arithmetic operations, and generate program-like reasoning chains to derive answers. While FinQA advances domain-specific QA, it struggles with complex tables and implicit numerical relationships requiring contextual reasoning beyond arithmetic. Furthermore, its narrow focus on structured numerical QA tasks and the dependence on predefined report structures limit its adaptability to evolving real-world financial tasks. FinBen (Xie et al., 2024) is a benchmark for financial reasoning that integrates numerical analysis, textual comprehension, and multi-modal data from financial reports, emphasizing tasks like ratio computation, trend prediction, and decision making. However, FinBen overemphasizes on quantitative tasks and underrepresents qualitative reasoning. And its reliance on idealized document formats ignores real-world noises and limits its generalizability. FinTSB (Hu et al., 2025) focuses on time-series forecasting with high-frequency trading data but overlooks exogenous factors. Other related financial benchmarks such as FinanceBench (Islam et al., 2023), PIXIU (Xie et al., 2023), and BizBench (Koncel-Kedziorski et al., 2023), offer limited evaluation task diversity and emphasize NLP capabilities, e.g., information extraction and QA, overlooking complex financial reasoning or practical application scenarios. SECQUE (Yoash et al., 2025) is a novel benchmark that advances the evaluation of LLMs in finance by simulating real-world financial challenges. It focuses on practical financial tasks and requires multi-step reasoning to handle noisy data akin to real-world scenarios. However, SECQUE still relies on a static dataset and may not fully capture the dynamic reality of financial workflows.

Financial LLMs and Agents. Recent advancements in LLMs have spurred the development of domain-specific LLMs and agents. FinGPT (Liu et al., 2023) is an open-sourced and data-centric framework, which uses real-time market data and leverages techniques such as Reinforcement Learning with Stock Prices (RLSP) to help models adapt to financial trends. Concurrently, FinAgent (Zhang et al., 2024b) introduces a multimodal, agent-based system enhanced with tools, e.g., data retrieval mechanism and chain-of-thought (COT) reasoning, for diverse financial trading tasks. FinCon (Yu et al., 2024) is an LLM-based multi-agent system designed for complex financial decision-making tasks such as stock trading and portfolio management. It employs a hierarchical manager-analyst framework inspired by real-world investment firms, enabling synchronized agent collaboration through natural language. Baichuan4-Finance (Zhang et al., 2024a) is a specialized LLM optimized for financial applications, which is built upon Baichuan's general AI capabilities and is fine-tuned with extensive financial data to enhance performance in financial tasks, e.g., financial

analysis, risk assessment and market prediction. DianJin-R1 (Zhu et al., 2025) is a financial LLM enhanced through Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a reinforcement learning method that incorporates dual reward signals, i.e., format reward and accuracy reward, guiding the model to excel in complex financial reasoning tasks.

C Preliminaries of Finance Management Workflows

Accounting (Godfrey et al., 2010) is the systematic practice of recording, summarizing, analyzing, and reporting financial transactions to ensure transparency, compliance, and informed decision-making. It is an essential component of business growth and sustainability. The main purpose of the program is to prepare and disseminate financial reports, speak from the essence, and to fundamentally relate to recording, reporting and resolving financial transactions to support reasonable decisions (Santos et al., 2020). After completing daily business operations, financial activities are recorded and classified as relevant accounts and adjusted to comply with the principles of the meeting to ensure that income and fees are matched over the corresponding meeting period. Finally, the adjusted account balance provides basic data for preparing comprehensive financial reports.

Auditing refers to the systematic and independent examination and evaluation of an organization's financial statements, operational processes, and regulatory compliance conducted by internal or external auditing entities or personnel. In contemporary practice, the scope of review has exceeded the scope of traditional financial reports. Nowadays, people are paying more and more attention to the review and analysis of data generated by the organization's daily operations. This evolution reflects the shift of audits to a more holistic approach, and modern audits not only emphasize the evaluation of financial reporting, but also use advanced data analysis to evaluate daily operational processes and internal controls (Ovami & Muda, 2023).

Consulting refers to professional services that help organizations solve problems and achieve objectives through systematic analysis (Biech, 2019). Clients typically seek consulting support to improve business performance or address operational challenges (Bruhn et al., 2018). The core of consulting services resides in diagnostic analysis of client operational status, with financial diagnostics emerging as the most strategically critical analytical dimension. This process involves deconstructing financial statements to build a quantitative evaluation framework. Key metrics include profitability (gross/net margins), operational efficiency (inventory/receivables turnover), and solvency (current/quick ratios). Leveraging established analytical paradigms such as DuPont Analysis (Soliman, 2008) and Altman Z-score models (Altman et al., 2017), this financial diagnostic methodology achieves dual objectives: i) precise identification of resource allocation inefficiencies in corporate operations, and ii) revelation of competitive positioning within industry landscapes through benchmarking analysis against peer comparables. These insights enable data-driven decisions for strategic restructuring, cost optimization, and capital allocation. Financial analytics not only differentiates consultants from domain experts but also validates the feasibility of cross-disciplinary solutions. Thus, financial analysis acts as both a strategic foundation for consulting and a bridge connecting financial data to business realities.

BALANCE SHEET

Assets	**Initial Amount**	**End Amount**
Current Assets		
Cash on Hand	3000000	270005.9
Bank Deposits	5000000	164645.57
Interest Receivable	0	2672.87
Accounts Receivable	0	2429482.13
Inventory	0	5090000
Total Current Assets	8000000	7956806.47
Non-Current Assets		
Fixed Assets	5000000	5305354.43
Accumulated Depreciation	0	(45751.41)
Net Fixed Assets	5000000	5259603.02
Total Non-Current Assets	5000000	5259603.02
Total Assets	13000000	13216409.49
Liabilities		
Current Liabilities		
Accounts Payable	0	1590000
Taxes Payable	0	271550.92
Total Current Liabilities	0	1861550.92
Total Liabilities	0	1861550.92
Owner's Equity		
Paid-in Capital	13000000	13000000
Retained Earnings	0	-1645141.46
Total Owner's Equity	13000000	11354859
Total Liabilities and Equity	13000000	13216409

Table 5: Balance sheet is a financial status report for a company at a specific time, reflecting all assets, debts and shareholder rights owned by the company

INCOME STATEMENT

Revenue	
Main Business Revenue	5431018.59
Total Revenue	5431018.59
Cost	
Cost of Goods Sold	(4410000)
Total Cost	(4410000)
Gross Profit	1021018.59
Expense	
Administrative Expenses	(1425164.2)
Selling Expenses	(493854.67)
Depreciation	(45751.41)
Financial Expenses	(432511.69)
Total Expenses	(2397281.97)
Other Revenue	
Interest Income	2672.87
Profit Before Tax	-1373590.51
Tax Expense	271550.92
Net Profit	-1645141.43

Table 6: Income statement is a financial report that shows a company's revenues, costs, and expenses over a period of time, reflecting the company's profitability and performance

CASH FLOW STATEMENT

Cash Flows from Operating Activities	
Net profit	-1645141.43
Depreciation	45751.41
(Increase) Decrease in Current Assets	
Accounts Receivable	(2429482.13)
Interest Receivable	(2672.87)
Inventory	(5090000))
Total (Increase) Decrease in Current Assets	(7522155)
Increase (Decrease) in Current Liabilities	
Accounts Payable	1590000
Tax Payable	271550.92
Total Increase (Decrease) in Current Liabilities	1861550.92
Net Cash Flow From Operations	-7259994.1
Cash Flows from Investing Activities	
Purchase of Fixed Assets	305354.43
Net Cash Flows from Investing Activities	(305354.43)
Beginning Cash and Cash Equivalents Balance	8000000
Ending Cash and Cash Equivalents Balance	434651.47
Net Increase	(7565348.53)

Table 7: Cash flow statement is a financial report that shows a company's cash inflows and outflows over a period of time, reflecting how the company generates and uses its cash through operating, investing, and financing activities

D FinSim

D.1 Types of Companies

The configurations of different types of companies are displayed in Table **??**, and the comparison between the companies is shown in Figure 6. Specifically, for initial capital, we use the sum of the initial bank deposit, the initial fixed assets, and the purchase unit price to represent; for features including profit margin and different frequencies, we display both minimum and maximum values.

- Type I considers capital goods manufacturers, e.g., heavy machinery and shipbuilding companies, which are characterized by capital-intensive operations with low sales frequency but premium purchase unit prices, reflecting specialized, high-value products.
- Type II considers transaction-driven companies, e.g., Chemical trader and industrial product distributor. These companies often face stable procurement costs but lack pricing power, leading to low, volatile gross margins. To maintain sales and market share, they rely on bulk purchasing and large-scale sales, despite high selling and administrative costs.
- Type III considers companies that offer high value-added consumer goods, such as luxury brands or premium electronics manufacturers. These businesses are characterized by high gross margins and low production costs. To sustain high revenue levels, they often make significant investments in selling expenses, particularly in branding and marketing efforts.
- Type IV considers asset-light companies, e.g., consulting and designing companies, which operate light-asset models with minimal fixed assets, high profit margin, and even no inventory. These businesses typically have a high purchase-on-credit rate, relying on credit for procurement, while maintaining robust profitability due to their low capital requirements and high-margin service models.
- Type V considers high-turnover companies, e.g., hotel and catering enterprises, which are characterized by high sales frequency, low unit prices, large quantity per purchase, and a dispersed customer base.

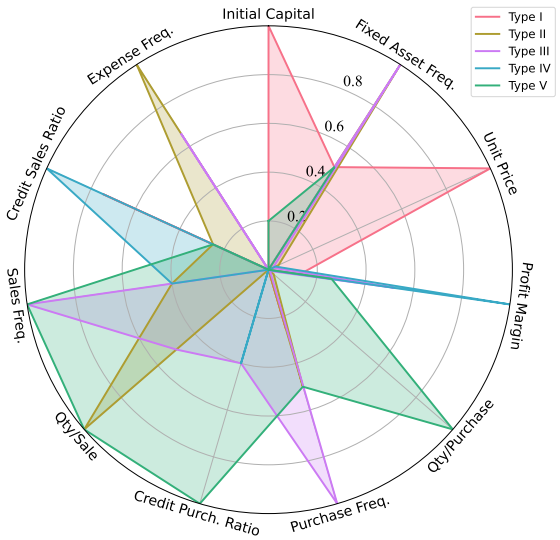


Figure 6: Companies Comparison.

E FinSuite

In this section, we present the complete information of the tasks for financial literacy, accounting, auditing and consulting considered in **FinMaster**, detailing each task's name, difficulty, description, and input and output specifications.

E.1 Financial Statement Items Definition

Item Name	**Item Definition**
Cash on Hand	Cash held by an entity that is available for use in its day-to-day operations.
Bank Deposits	Bank deposits are funds deposited into a bank or other financial institution.
Interest Receivable	Amounts of interest accrued but not yet received.
Accounts Receivable	Amounts owed to the entity for goods or services sold or provided on credit.
Inventory	Assets held for sale in the ordinary course of business, in production for such sale, or in the process of being manufactured.
Total Current Assets	The total amount of assets that are expected to be realised or intended for sale or consumption in the normal course of the entity's operating cycle.
Fixed Assets	Tangible items that are held for use in the production or supply of goods or services, for rental to others, or for administrative purposes.
Accumulated Depreciation	The total amount of depreciation recognised as an expense in the statement of profit or loss and other comprehensive income.
Total Non-current Assets	The total amount of assets that are not expected to be realised or intended for sale or consumption in the normal course of the entity's operating cycle.
Total Assets	The total present economic resources controlled by the entity as a result of past events, which also means the sum of all assets owned by an entity, both current and non-current, that are expected to bring future economic benefits to the company.
Accounts Payable	Amounts owed by the entity for goods or services received or purchased on credit.
Taxes Payable	Amounts of taxes accrued but not yet paid.
Total Current Liabilities	The total amount of liabilities that are expected to be settled in the normal course of the entity's operating cycle.
Paid-in Capital	The amount of capital contributed by shareholders in exchange for shares.
Retained Earnings	The amount of profit or loss retained in the entity, rather than being distributed to shareholders.
Total Owner's Equity	The total amount of equity recognised in the statement of financial position.
Total Liabilities and Owner's Equity	The total amount of liabilities and equity recognised in the statement of financial position.

Table 8: Definition of Balance Sheet Items

Item Name	**Item Definition**
Main Business Revenue	Income arising in the course of the entity's core operating activities.
Total Revenue	Total income arising in the course of an entity's ordinary activities.
Cost of Goods Sold	Carrying amount of inventories sold during the reporting period.
Total Cost	The aggregate of all expenses incurred by a company to generate its revenues during a specific accounting period.
Gross Profit	Gross profit is the difference between sales revenue and the cost of goods sold. Gross profit is the cleanest accounting measure of true economic profitability.
Administrative Expenses	The costs of distribution or administrative activities; costs of general management and administration of the entity as a whole.
Selling Expenses	Costs incurred to secure customer orders and to deliver the goods and services to customers.
Depreciation	The systematic allocation of the depreciable amount of an asset over its useful life.
Financial Expenses	Financing costs incurred by an enterprise to raise funds needed for production and operation.
Total Expenses	The total amount of expenses incurred by an entity during a reporting period.
Interest Income	Income earned by an entity from financial assets.
Profit Before Tax	Profit or loss for a period before deducting tax expense. It represents the company's earnings from all activities—operating and non-operating—before the effects of tax expenses.
Tax Expense	Total amount of taxes an entity is expected to pay or recover during a reporting period.
Net Profit	The amount of profit an entity retains after all expenses, including operating costs, interest, taxes, depreciation, and amortization, have been deducted from total revenue.

Table 9: Definition of Income Statement Items

Item Name	**Item Definition**
Cash Flow From Operating Activities	The cash inflows and outflows generated by a company's core business operations during a specific period.
Net Profit	The amount of profit an entity retains after all expenses, including operating costs, interest, taxes, depreciation, and amortization, have been deducted from total revenue.
Depreciation	The systematic allocation of the depreciable amount of an asset over its useful life.
(Increase) Decrease in Accounts Receivable	The reduction in the amounts owed to the entity for goods or services sold or provided on credit during the period.
(Increase) Decrease in Interest Receivable	The reduction in the amount of interest accrued but not yet received during the period.
(Increase) Decrease in Inventory	The reduction in the amount of assets held for sale in the ordinary course of business, in production for such sale, or in the process of being manufactured during the period.
Increase (Decrease) in Accounts Payable	The addition in the amount owed by the entity for goods or services received or purchased on credit during the period.
Increase (Decrease) in Tax Payable	The addition in the amount of taxes accrued but not yet paid during the period.
Net Cash Flow from Operating Activities	The total cash generated or used by a company's core business operations after accounting for all cash inflows and outflows within a specific period.
Cash Flow from Investing Activities	Investing activities are the acquisition and disposal of long-term assets and other investments not included in cash equivalents and the receipt of interest and dividends.
Purchase of Fixed Assets	The acquisition of property, plant and equipment.
Net Cash Flow from Investing Activities	The net amount of cash and cash equivalents generated from an entity's activities that are the acquisition and disposal of long-term assets and other investments not included in cash equivalents and the receipt of interest and dividends.
Beginning Balance	The amount of cash and cash equivalents at the beginning of the period.
Ending Balance	The amount of cash and cash equivalents at the end of the period.
Net Increase	The net addition in the amount of cash and cash equivalents during the period.

Table 10: Definition of Cash Flow Statement Items

E.2 Element Categorization

Task Name	Error Category
Transaction TYPE Record Error	Record Error
Transaction DATE Record Error	Record Error
Transaction PAYMENT/RECEIPT_STATUS Record Error	Record Error
Transaction PAYMENT_METHOD Record Error	Record Error
Transaction QUANTITY Record Error	Record Error
Transaction UNIT_PRICE Record Error	Record Error
Transaction RECEIVE_METHOD Record Error	Record Error
Transaction AMOUNT Calculation Error	Calculation Error
Transaction TAX_AMOUNT Calculation Error	Calculation Error
Transaction PROFIT Calculation Error	Calculation Error
Transaction Without PREPARER Error	Transaction Approval Mismatch
Transaction Without APPROVER Error	Transaction Approval Mismatch

Table 11: Audit Basic Error Classification

Indicator	Category
Free Cash Flow (FCF)	Cash Flow Quality
Operating Cash Flow to Net Income Ratio	Cash Flow Quality
Operating Cash Flow Ratio	Cash Flow Quality
Gross Profit Margin	Profitability
Net Profit Margin	Profitability
Return on Assets (ROA)	Profitability
Return on Equity (ROE)	Profitability
Current Ratio	Liquidity
Quick Ratio	Liquidity
Cash to Current Debt Ratio	Liquidity
Operating Cash Flow to Current Liabilities Ratio	Liquidity
Debt to Asset Ratio	Solvency
Debt to Equity Ratio	Solvency
Cash Flow to Debt Ratio	Solvency
Inventory Turnover Ratio	Operational Efficiency
Accounts Receivable Turnover Ratio	Operational Efficiency
Current Assets Turnover Ratio	Operational Efficiency
Total Asset Turnover Ratio	Operational Efficiency

Table 12: Financial Indicators Classification

E.3 Critical Financial Indicators Display

Indicator	Description	Formula
Free Cash Flow (FCF)	The cash remaining after a company has paid for its operating expenses and capital expenditures.	Net Cash Flow from Operating Activities − Purchase of Fixed Assets
Operating Cash Flow to Net Income Ratio	A ratio that evaluates the relationship between cash generated from operating activities and net income.	Net Cash Flow from Operating Activities/Net Profit
Operating Cash Flow Ratio	A liquidity metric that measures the adequacy of operating cash flow in covering a company's short-term liabilities.	Net Cash Flow from Operating Activities/Current Liabilities
Gross Profit Margin	A profitability ratio calculated as gross profit divided by revenue, expressed as a percentage.	$((\text{Revenue} - \text{COGS})/\text{Revenue}) \times 100\%$
Net Profit Margin	A financial metric that shows the percentage of net income derived from total revenue.	$(\text{Net Profit}/\text{Revenue}) \times 100\%$
Return on Assets (ROA)	A profitability ratio that measures the efficiency with which a company utilizes its total assets to generate net income.	$((2 * \text{Net Profit})/(\text{Beginning Total Assets} + \text{Ending Total Assets})) \times 100\%$
Return on Equity (ROE)	A performance metric that quantifies the return generated on shareholders' equity.	$((2 * \text{Net Profit})/(\text{Beginning Owner's Equity} + \text{Ending Owner's Equity})) \times 100\%$
Current Ratio	A liquidity ratio calculated as current assets divided by current liabilities.	Current Assets/Current Liabilities
Quick Ratio	A stringent liquidity measure that assesses a company's ability to pay off its current liabilities using its most liquid assets.	(Current Assets − Inventory)/Current Liabilities
Cash to Current Debt Ratio	A liquidity ratio that evaluates the proportion of cash and cash equivalents available to settle current liabilities, indicating short-term financial stability.	(Cash and Cash Equivalents – Ending Balance)/Current Liabilities
Operating Cash Flow to Current Liabilities Ratio	A ratio that measures the sufficiency of cash generated from operating activities to cover current liabilities, reflecting operational efficiency and liquidity.	Net Cash Flow from Operating Activities/Ending Current Liabilities
Debt to Asset Ratio	A leverage ratio that calculates the percentage of a company's total assets financed through debt. It is determined by dividing total debt by total assets.	Total Liabilities/Total Assets
Debt to Equity Ratio	A financial leverage metric that compares a company's total debt to its shareholders' equity, illustrating the proportion of debt used relative to equity financing.	Total Liabilities/Owner's Equity
Cash Flow to Debt Ratio	A solvency ratio that measures a company's ability to repay its total debt using cash generated from operating activities.	Net Cash Flows from Operating Activities/Total Liabilities
Inventory Turnover Ratio	An efficiency metric that calculates how many times a company sells and replaces its inventory over a specific period.	$(2 * \text{COGS})/(\text{Beginning Inventory} + \text{Ending Inventory})$
Accounts Receivable Turnover Ratio	A ratio that measures the efficiency of a company in collecting its accounts receivable.	$(2 * \text{Revenue})/(\text{Beginning Accounts Receivable} + \text{Ending Accounts Receivable})$
Current Assets Turnover Ratio	An efficiency ratio that evaluates how effectively a company utilizes its current assets to generate revenue.	$(2 * \text{Revenue})/(\text{Beginning Current Assets} + \text{Ending Current Assets})$
Total Asset Turnover Ratio	A financial efficiency metric that measures the ability of a company to generate revenue from its total assets.	$(2 * \text{Revenue})/(\text{Beginning Total Assets} + \text{Ending Total Assets})$

Table 13: Critical Financial Indicators Description and Formula

E.4 Financial Statement Tasks Information

Task Name	{α,β,γ}	Task Description	Input	Output
Financial Literacy Detection-Cash on Hand	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of cash held by an entity that is available for use in its day-to-day operations, including initial and final value	Balance sheet	The value of cash on hand, including initial and final value
Financial Literacy Detection-Bank Deposits	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of funds deposit into a bank, including initial and final value	Balance sheet	The value of bank deposits, including initial and final value
Financial Literacy Detection-Accounts Receivable	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of amounts owed to the entity for goods or services sold or provided on credit, including initial and final value	Balance sheet	The value of accounts receivable, including initial and final value
Financial Literacy Detection-Interest Receivable	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of amounts of interest accrued but not yet received, including initial and final value	Balance sheet	The value of interest receivable, including initial and final value
Financial Literacy Detection-Inventory	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of assets held for sale in the ordinary course of business, in production for such sale, or in the process of being manufactured, including initial and final value	Balance sheet	The value of inventory, including initial and final value
Financial Literacy Detection-Fixed Assets	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of tangible items that are held for use in the production or supply of goods or services, for rental to others, or for administrative purposes, including initial and final value	Balance sheet	The value of fixed assets, including initial and final value
Financial Literacy Detection-Accumulated Depreciation	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of expense in the statement of profit or loss and other comprehensive income, including initial and final value	Balance sheet	The value of accumulated depreciation, including initial and final value
Financial Literacy Detection-Accounts Payable	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of the amounts owed by the entity for goods or services received or purchased on credit, including initial and final value	Balance sheet	The value of accounts payable, including initial and final value
Financial Literacy Detection-Taxes Payable	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of taxes accrued but not yet paid, including initial and final value	Balance sheet	The value of taxes payable, including initial and final value
Financial Literacy Detection-Paid-in Capital	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of capital contributed by shareholders in exchange for shares, including initial and final value	Balance sheet	The value of paid-in capital, including initial and final value
Financial Literacy Detection-Retained Earnings	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of profit or loss retained in the entity, rather than being distributed to shareholders, including initial and final value	Balance sheet	The value of retained earnings, including initial and final value
Financial Literacy Detection-Current Assets	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of total assets that are expected to be realised or intended for sale or consumption in the normal course of the entity's operating cycle, including initial and final value	Balance sheet	The value of current assets, including initial and final value
Financial Literacy Detection-Non-current Assets	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of total assets that are not expected to be realised or intended for sale or consumption in the normal course of the entity's operating cycle, including initial and final value	Balance sheet	The value of non-current assets, including initial and final value
Financial Literacy Detection-Current Liabilities	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of total amount of liabilities that are expected to be settled in the normal course of the entity's operating cycle, including initial and final value	Balance sheet	The value of current liabilities, including initial and final value
Financial Literacy Detection-Owner's Equity	{2,1,2}	Based on the balance sheet, identify and extract the specific line items and value of total amount of equity recognised in the statement of financial position, including initial and final value	Balance sheet	The value of owner's equity, including initial and final value
Financial Literacy Detection-Total Liabilities and Owner's Equity	{5,1,5}	Based on the balance sheet, identify and extract the specific line items and value of the total amount of liabilities and equity recognised in the statement of financial position, along with the relevant financial data involved in the calculation, including initial and final value. In addition, decompose this item into 2 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Balance sheet	i) The value of liabilities and owner's equity, including initial and final value ii) The value of each core sub-item under owner's equity, including initial and final value
Financial Literacy Detection-Accounts Receivable & Accounts Payable	{2,1,2}	Based on the balance sheet, identify and extract the two specific line items and value of amounts owed to the entity for goods or services sold or provided on credit and the amounts owed by the entity for goods or services received or purchased on credit, including initial and final value. For multiple outputs, maintain the original line item order as shown in the input statement.	Balance sheet	The value of accounts receivable and accounts payable, including initial and final value
Financial Literacy Detection-Cash on Hand & Fixed Assets & Taxes Payable	{3,1,3}	Based on the balance sheet, identify and extract the three specific line items and value of cash held by an entity that is available for use in its day-to-day operations, tangible items that are held for use in the production or supply of goods or services, for rental to others, or for administrative purposes and the amounts of taxes accrued but not yet paid, including initial and final value. For multiple outputs, maintain the original line item order as shown in the input statement.	Balance sheet	The value of cash on hand, fixed assets and taxes payable, including initial and final value
Financial Literacy Detection-Interest Receivable & Accumulated Depreciation & Taxes Payable & Paid-in Capital	{4,1,4}	Based on the balance sheet, identify and extract the four specific line items and value of amounts of interest accrued but not yet received, accumulated the systematic allocation of the depreciable amount of an asset over its useful life, tax payable and capital contributed by shareholders in exchange for shares, including initial and final value. For multiple outputs, maintain the original line item order as shown in the input statement.	Balance sheet	The value of interest receivable, accumulated depreciation, tax payable and paid-in capital, including initial and final value

Table 14: Task Information Table - Balance Sheet Detection in Financial Literacy

Task Name	{α,β,γ}	Task Description	Input	Output
Financial Literacy Detection-Cost of Goods Sold	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of carrying amount of inventories sold during the reporting period	Income statement	The value of cost of goods sold
Financial Literacy Detection-Main Business Revenue	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of income arising in the course of the entity's core operating activities	Income statement	The value of main business revenue
Financial Literacy Detection-Gross Profit	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of difference between revenue and cost	Income statement	The value of gross profit
Financial Literacy Detection-Interest Income	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of income earned by an entity from financial assets	Income statement	The value of interest income
Financial Literacy Detection-Administrative Expenses	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of costs of general management and administration of the entity as a whole	Income statement	The value of administrative expenses
Financial Literacy Detection-Selling Expenses	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of costs incurred to secure customer orders and to deliver the goods and services to customers	Income statement	The value of selling expenses
Financial Literacy Detection-Financial Expenses	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of financing costs incurred by an enterprise to raise funds needed for production and operation	Income statement	The value of financial expenses
Financial Literacy Detection-Accumulated Depreciation	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of accumulated the systematic allocation of the depreciable amount of an asset over its useful life	Income statement	The value of accumulated depreciation
Financial Literacy Detection-Tax Expense	{1,1,1}	Based on the income statement, identify and extract the specific line items and value of total amount of taxes an entity is expected to pay or recover during a reporting period	Income statement	The value of tax expense
Financial Literacy Detection-Total Revenue	{2,1,2}	Based on the income statement, identify and extract the specific line items and value of total income arising in the course of an entity's ordinary activities, along with the values and names of its constituent line items. In addition, decompose this item into 1 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Income statement	i) The value of revenue ii) The value of each sub-item under revenue
Financial Literacy Detection-Total Expenses	{5,1,5}	Based on the income statement, identify and extract the specific line items and value of operating expenses, along with the relevant financial data involved in the calculation. In addition, decompose this item into 4 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Income statement	i) The value of total expenses ii) The value of each sub-item under operating expenses
Financial Literacy Detection-Profit Before Tax	{5,1,5}	Based on the income statement, identify and extract the specific line items and value of profit or loss for a period before deducting tax expense, along with the relevant financial data involved in the calculation. In addition, decompose this item into 4 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Income statement	i) The value of profit before tax ii) The value of each sub-item under profit before tax
Financial Literacy Detection-Net Profit	{6,1,6}	Based on the income statement, identify and extract the specific line items and value of the amount of profit an entity retains after all expenses, including operating costs, interest, taxes, depreciation, and amortization, have been deducted from total revenue, along with the relevant financial data involved in the calculation. In addition, decompose this item into 5 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Income statement	i) The value of net profit ii) The value of core sub-item under operating expenses
Financial Literacy Detection-Main Business Revenue & Cost of Goods Sold	{2,1,2}	Based on the income statement, identify and extract the two specific line items and value of income arising in the course of the entity's core operating activities and carrying amount of inventories sold during the reporting period. For multiple outputs, maintain the original line item order as shown in the input statement.	Income statement	The value of main business revenue and cost of goods sold
Financial Literacy Detection-Total Revenue & Cost of Goods Sold & Administrative Expenses	{3,1,3}	Based on the income statement, identify and extract the three specific line items and value of total income arising in the course of an entity's ordinary activities, carrying amount of inventories sold during the reporting period, and costs of general management and administration of the entity as a whole. For multiple outputs, maintain the original line item order as shown in the input statement.	Income statement	The value of revenue, cost of goods sold and administrative expenses
Financial Literacy Detection-Selling Expenses & Depreciation & Financial Expenses & Tax Expense	{4,1,4}	Based on the income statement, identify and extract the four specific line items and value of financing costs incurred by an enterprise to raise funds needed for production and operation, the systematic allocation of the depreciable amount of an asset over its useful life, and total amount of taxes an entity is expected to pay or recover during a reporting period. For multiple outputs, maintain the original line item order as shown in the input statement.	Income statement	The value of financial expenses, selling expenses, depreciation and tax expense

Table 15: Task Information Table - Income Statement Detection in Financial Literacy

Task Name	{α,β,γ}	Task Description	Input	Output
Financial Literacy Detection-Net Profit	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of the amount of profit an entity retains after all expenses, including operating costs, interest, taxes, depreciation, and amortization, have been deducted from total revenue	Cash flow statement	The value of net profit
Financial Literacy Detection-Depreciation	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of the systematic allocation of the depreciable amount of an asset over its useful life	Cash flow statement	The value of depreciation
Financial Literacy Detection-Decrease in Accounts Receivable	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of decrease in amounts owed to the entity for goods or services sold or provided on credit	Cash flow statement	The value of decrease in accounts receivable
Financial Literacy Detection-Decrease in Inventory	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of reduction in the amount of assets held for sale in the ordinary course of business, in production for such sale, or in the process of being manufactured during the period	Cash flow statement	The value of decrease in inventory
Financial Literacy Detection-Increase in Accounts Payable	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of the addition in the amount owed by the entity for goods or services received or purchased on credit during the period	Cash flow statement	The value of increase in accounts payable
Financial Literacy Detection-Increase in Taxes Payable	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of increase in the amounts of taxes accrued but not yet paid	Cash flow statement	The value of increase in taxes payable
Financial Literacy Detection-Purchase of Fixed Assets	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of cash payments to acquire property, plant and equipment and other long-term assets	Cash flow statement	The value of purchased of fixed assets
Financial Literacy Detection-Beginning Cash and Cash Equivalents Balance	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of cash and cash equivalents at the beginning of the period	Cash flow statement	The value of Beginning Cash and Cash Equivalents Balance
Financial Literacy Detection-Ending Cash and Cash Equivalents Balance	{1,1,1}	Based on the cash flow statement, identify and extract the specific line items and value of cash and cash equivalents at the end of the period	Cash flow statement	The value of Ending Cash and Cash Equivalents Balance
Financial Literacy Detection-Net Cash Flow from Operating Activities	{7,1,7}	Based on the net cash flow statement, identify and extract the specific line items and value of net amount of cash and cash equivalents generated from an entity's activities that are the principal revenue-producing activities of the entity and other activities that are not investing or financing activities, along with the relevant financial data involved in the calculation. In addition, decompose this item into 7 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Cash flow statement	i) The value of net cash flow from operating activities ii) The value of core sub-item under net cash flow from operating activities
Financial Literacy Detection-Net Cash Flow from Investing Activities	{2,1,2}	Based on the cash flow statement, identify and extract the specific line items and value of net amount of cash and cash equivalents generated from an entity's activities that are the acquisition and disposal of long-term assets and other investments not included in cash equivalents and the receipt of interest and dividends, along with the relevant financial data involved in the calculation. In addition, decompose this item into 1 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Cash flow statement	i) The value of net cash flow from investing activities ii) The value of core sub-item under net cash flow from operating activities
Financial Literacy Detection-Net Increase in Cash and Cash Equivalents	{3,1,3}	Based on the cash flow statement, identify and extract the specific line items and value of net addition in the amount of cash and cash equivalents during the period, along with the relevant financial data involved in the calculation. In addition, decompose this item into 2 component sub-items, all of which must also originate from the input statement. For each sub-item, output its initial and final values.	Cash flow statement	i) The value of cash flow from net increase in cash and cash equivalents ii) The value of core sub-item under net cash flow from operating activities
Financial Literacy Detection-Net Profit & Purchase of Fixed Assets	{2,1,2}	Based on the cash flow statement, identify and extract the two specific line items and value of the amount of profit an entity retains after all expenses, including operating costs, interest, taxes, depreciation, and amortization, have been deducted from total revenue and cash payments to acquire property, plant and equipment and other long-term assets. For multiple outputs, maintain the original line item order as shown in the input statement.	Cash flow statement	The value of net profit and purchased in fixed assets
Financial Literacy Detection-Increase in Accounts Payable & Purchase of Fixed Assets & Beginning Cash Balance	{3,1,3}	Based on the cash flow statement, identify and extract the three specific line items and value of the addition in the amount owed by the entity for goods or services received or purchased on credit during the period, acquisition of property, plant and equipment, and cash and cash equivalents at the beginning of the period. For multiple outputs, maintain the original line item order as shown in the input statement.	Cash flow statement	The value of increase in accounts payable, purchase of fixed assets and beginning cash and cash equivalents balance
Financial Literacy Detection-Depreciation & Decrease in Inventory & Net Cash Flow from Investing & Net Increase	{4,1,4}	Based on the cash flow statement, identify and extract the four specific line items and value of the systematic allocation of the depreciable amount of an asset over its useful life, reduction in the amount of assets held for sale in the ordinary course of business, in production for such sale, or in the process of being manufactured during the period, net amount of cash and cash equivalents generated from an entity's activities that are the acquisition and disposal of long-term assets and other investments not included in cash equivalents and the receipt of interest and dividends, and net addition in the amount of cash and cash equivalents during the period. For multiple outputs, maintain the original line item order as shown in the input statement.	Cash flow statement	The value of depreciation, decrease in inventory, net cash flow from investing activities and net increase

Table 16: Task Information Table - Cash Flow Statement Detection in Financial Literacy

Task Name	{α,β,γ}	Task Description	Input	Output
Financial Literacy Detection-Interest Receivable	{1,3,1}	Based on the all financial statements, identify and extract the specific line items and value of amounts of interest accrued but not yet received, including initial and final value	All financial statements	The value of interest receivable, including initial and final value
Financial Literacy Detection-Paid-in Capital	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of capital contributed by shareholders in exchange for shares, including initial and final value	All financial statements	The value of paid-in capital, including initial and final value
Financial Literacy Detection-Cost of Goods Sold	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of carrying amount of inventories sold during the reporting period	All financial statements	The value of cost of goods sold
Financial Literacy Detection-Selling Expenses	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of selling expenses	All financial statements	The value of selling expenses
Financial Literacy Detection-Tax Expense	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of total amount of taxes an entity is expected to pay or recover during a reporting period	All financial statements	The value of tax expense
Financial Literacy Detection-Depreciation	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of the systematic allocation of the depreciable amount of an asset over its useful life	All financial statements	The value of depreciation
Financial Literacy Detection-Increase in Accounts Payable	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of the addition in the amount owed by the entity for goods or services received or purchased on credit during the period	All financial statements	The value of increase in accounts payable
Financial Literacy Detection-Beginning Cash and Cash Equivalents Balance	{1,3,1}	Based on the financial statement, identify and extract the specific line items and value of cash and cash equivalents at the beginning of the period	All financial statements	The value of Beginning Cash and Cash Equivalents Balance
Financial Literacy Detection-Interest Receivable & Net Increase in Cash	{2,3,2}	Based on the all financial statements, identify and extract the two specific line items and end value of amounts of interest accrued but not yet received, and net addition in the amount of cash and cash equivalents during the period. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The end value of interest receivable and net increase of cash and cash equivalents
Financial Literacy Detection-Bank Deposits & Interest Income	{2,3,2}	Based on the all financial statements, identify and extract the two specific line items and end value of funds deposit into a bank, and income earned by an entity from financial assets. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The end value of bank deposits and interest income
Financial Literacy Detection-Selling Expenses & Purchase of Fixed Assets	{2,3,2}	Based on the all financial statements, identify and extract the two specific line items and value of costs incurred to secure customer orders and to deliver the goods and services to customers, and cash payments to acquire property, plant and equipment and other long-term assets. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The value of selling expenses and purchased of fixed assets
Financial Literacy Detection-Accounts Receivable & Financial Expenses & Fixed Assets	{3,3,3}	Based on the all financial statements, identify and extract the three specific line items and end value of amounts owed to the entity for goods or services sold or provided on credit, financing costs incurred by an enterprise to raise funds needed for production and operation, and cash payments to acquire property, plant and equipment and other long-term assets. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The end value of accounts receivable, financial expenses and purchased of fixed assets
Financial Literacy Detection-Taxes Payable & Revenue & Operating Cash Flow	{3,3,3}	Based on the all financial statements, identify and extract the three specific line items and end value of net amount of cash and cash equivalents generated from an entity's activities that are the principal revenue-producing activities of the entity and other activities that are not investing or financing activities, the amounts of taxes accrued but not yet paid, and total income arising in the course of an entity's ordinary activities. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The end value of net cash flow from operating activities, taxes payable and revenue
Financial Literacy Detection-Paid-in Capital & Profit Before Tax & Accounts Payable	{3,3,3}	Based on the all financial statements, identify and extract the three specific line items and end value of profit or loss for a period before deducting tax expense, the addition in the amount owed by the entity for goods or services received or purchased on credit during the period, and capital contributed by shareholders in exchange for shares. For multiple outputs, group them by financial statement in the following order: 1.Balance Sheet 2.Income Statement 3.Cash Flow Statement. Within each group, maintain the original line item order as shown in the input statement.	All financial statements	The end value of profit before tax, increase in accounts payable and paid-in capital

Table 17: Task Information Table - Financial Statement Detection in Financial Literacy

Task Name	{α,β,γ}	Task Description	Input	Output
Balance Sheet-Cash on Hand	{1,1,2}	Based on transactions data, calculate the total amount of cash on hand item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Cash on Hand, including initial and final value
Balance Sheet-Bank Deposits	{1,1,2}	Based on transactions data, calculate the bank deposits item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Bank Deposits, including initial and final value
Balance Sheet-Inventory	{1,1,2}	Based on transactions data, calculate the inventory item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Inventory, including initial and final value
Balance Sheet-Accounts Receivable	{1,1,2}	Based on transactions data, calculate the accounts receivable item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Accounts Receivable, including initial and final value
Balance Sheet-Interest Receivable	{1,1,2}	Based on transactions data, calculate the Interest Receivable item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Interest Receivable, including initial and final value
Balance Sheet-Current Assets	{5,1,2}	Based on transactions data, calculate the current assets item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Current assets, including initial and final value
Balance Sheet-Accumulated Depreciation	{1,1,2}	Based on transactions data, calculate the Accumulated Depreciation item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Accumulated Depreciation, including initial and final value
Balance Sheet-Fixed Assets net	{1,1,2}	Based on transactions data, calculate the Fixed Assets net item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Fixed Assets net, including initial and final value
Balance Sheet-Non-current Assets	{2,1,2}	Based on transactions data, calculate the property, plant and non-current assets item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Non-current Assets, including initial and final value
Balance Sheet-Total Assets	{7,1,2}	Based on transactions data, calculate the total assets item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Total Assets, including initial and final value
Balance Sheet-Accounts Payable	{1,1,2}	Based on transactions data, calculate the accounts payable item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Accounts Payable, including initial and final value
Balance Sheet-Taxes Payable	{1,1,2}	Based on transactions data, calculate the taxes payable item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Taxes Payable, including initial and final value
Balance Sheet-Current Liabilities	{2,1,2}	Based on transactions data, calculate the current liabilities item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Current Liabilities,including initial and final value
Balance Sheet-Total Liabilities	{2,1,2}	Based on transactions data, calculate the total liabilities item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Total Liabilities, including initial and final value
Balance Sheet-Paid-in Capital	{1,1,2}	Based on transactions data, calculate the Paid-in Capital item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Paid-in Capital, including initial and final value
Balance Sheet-Retained Earnings	{1,1,2}	Based on transactions data,calculate the retained earnings item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Retained Earnings, including initial and final value
Balance Sheet-Total Owner's Equity	{2,1,2}	Based on transactions data,calculate the total owner's equity item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Total Owner's Equity, including initial and final value
Balance Sheet-Total Liabilities and Owner's Equity	{4,1,2}	Based on transactions data,calculate the total liabilities and owner's equity item in the balance sheet, including both the initial and final amounts.	All transactions data	The total value of Total Liabilities and Owner's Equity, including initial and final value
Balance Sheet-Balance Sheet	{37,1,2}	Based on transactions data, directly generate a complete balance sheet, including both the initial and final amounts.	All transactions data	The complete balance sheet

Table 18: Task Information Table - Balance Sheet Generation in Accounting

Task Name	{α,β,γ}	Task Description	Input	Output
Income Statement-Main Business Revenue	{1,1,1}	Based on transactions data, calculate the final Main Business Revenue item in the Income.	Statement All transactions data	The final total value of Main Business Revenue
Income Statement-Total Revenue	{1,1,1}	Based on transactions data, calculate the final Total Revenue item in the Income Statement.	All transactions data	The final total value of Total Revenue
Income Statement-Cost of Goods Sold	{1,1,1}	Based on transactions data, calculate the final Cost of Goods Sold item in the Income Statement.	All transactions data	The final total value of Cost of Goods Sold
Income Statement-Total Cost	{1,1,1}	Based on transactions data, calculate the final Total Cost item in the Income Statement.	All transactions data	The final total value of Total Cost
Income Statement-Gross Profit	{2,1,1}	Based on transactions data, calculate the final Gross Profit item in the Income Statement.	All transactions data	The final total value of Gross Profit
Income Statement-Depreciation	{1,1,1}	Based on transactions data, calculate the final Depreciation item in the Income Statement.	All transactions data	The final total value of Depreciation
Income Statement-Administrative Expenses	{1,1,1}	Based on transactions data, calculate the final Administrative Expenses item in the Income Statement.	All transactions data	The final total value of Administrative Expenses
Income Statement-Sales Expenses	{1,1,1}	Based on transactions data, calculate the final Sales Expenses item in the Income Statement.	All transactions data	The final total value of Sales Expenses
Income Statement-Financial Expenses	{1,1,1}	Based on transactions data, calculate the final Financial Expenses item in the Income Statement.	All transactions data	The final total value of Financial Expenses
Income Statement-Total Expenses	{4,1,1}	Based on transactions data, calculate the final Total Expenses item in the Income Statement.	All transactions data	The final total value of Total Expenses
Income Statement-Interest Income	{1,1,1}	Based on transactions data, calculate the final Interest Income item in the Income Statement.	All transactions data	The final total value of Interest Income
Income Statement-Profit Before Tax	{7,1,1}	Based on transactions data, calculate the final Profit Before Tax item in the Income Statement.	All transactions data	The final total value of Profit Before Tax
Income Statement-Tax Expense	{1,1,1}	Based on transactions data, calculate the final Tax Expense item in the Income Statement.	All transactions data	The final total value of Tax Expense
Income Statement-Net Profit	{8,1,1}	Based on transactions data, calculate the final Net Profit item in the Income Statement.	All transactions data	The final total value of Net Profit
Income Statement-Income Statement	{31,1,1}	Based on transactions data, directly generate a complete income statement.	All transactions data	The complete income statement

Table 19: Task Information Table - Income Statement Generation in Accounting

Task Name	{α,β,γ}	Task Description	Input	Output
Cash Flow Statement-Net profit	{8,1,1}	Based on transactions data, calculate the final Net profit item in the Cash Flow Statement.	All transactions data	The final total value of Net profit
Cash Flow Statement-Depreciation	{1,1,1}	Based on transactions data, calculate the final Depreciation item in the Cash Flow Statement.	All transactions data	The final total value of Depreciation
Cash Flow Statement-Accounts Receivable	{1,1,1}	Based on transactions data, calculate the final Accounts Receivable item in the Cash Flow Statement.	All transactions data	The final total value of Accounts Receivable
Cash Flow Statement-Interest Receivable	{1,1,1}	Based on transactions data, calculate the final Interest Receivable item in the Cash Flow Statement.	All transactions data	The final total value of Interest Receivable
Cash Flow Statement-Inventory	{1,1,1}	Based on transactions data, calculate the final Inventory item in the Cash Flow Statement.	All transactions data	The final total value of Inventory
Cash Flow Statement-Total (Increase) Decrease in Current Assets	{1,1,1}	Based on transactions data, calculate the final Total (Increase) Decrease in Current Assets item in the Cash Flow Statement.	All transactions data	The final total value of Total (Increase) Decrease in Current Assets
Cash Flow Statement-Accounts Payable	{1,1,1}	Based on transactions data, calculate the final Accounts Payable item in the Cash Flow Statement.	All transactions data	The final total value of Accounts Payable
Cash Flow Statement-Tax Payable	{14,1,1}	Based on transactions data, calculate the final Tax Payable item in the Cash Flow Statement.	All transactions data	The final total value of Tax Payable
Cash Flow Statement-Total Increase (Decrease) in Current Liabilities	{1,1,1}	Based on transactions data, calculate the final Total Increase (Decrease) in Current Liabilities item in the Cash Flow Statement.	All transactions data	The final total value of Total Increase (Decrease) in Current Liabilities
Cash Flow Statement-Net Cash Flow from Operating Activities	{1,1,1}	Based on transactions data, calculate the final Net Cash Flow from Operating Activities item in the Cash Flow Statement.	All transactions data	The final total value of Net Cash Flow from Operating Activities
Cash Flow Statement-Purchase of Fixed Assets	{1,1,1}	Based on transactions data, calculate the final Purchase of Fixed Assets item in the Cash Flow Statement.	All transactions data	The final total value of Purchase of Fixed Assets
Cash Flow Statement-Net Cash Flows from Investing Activities	{1,1,1}	Based on transactions data, calculate the final Net Cash Flows from Investing Activities item in the Cash Flow Statement.	All transactions data	The final total value of Net Cash Flows from Investing Activities
Cash Flow Statement-Beginning Cash and Cash Equivalents Balance	{2,1,1}	Based on transactions data, calculate the final Beginning Cash and Cash Equivalents Balance item in the Cash Flow Statement.	All transactions data	The final total value of Beginning Cash and Cash Equivalents Balance
Cash Flow Statement-Ending Cash and Cash Equivalents Balance	{2,1,1}	Based on transactions data, calculate the final Ending Cash and Cash Equivalents Balance item in the Cash Flow Statement.	All transactions data	The final total value of Ending Cash and Cash Equivalents Balance
Cash Flow Statement-Net Increase	{4,1,1}	Based on transactions data, calculate the final Net Increase item in the Cash Flow Statement.	All transactions data	The final total value of Net Increase
Cash Flow Statement-Cash Flow Statement	{38,1,1}	Based on transactions data, directly generate a complete Cash Flow Statement.	All transactions data	The complete cash flow statement

Table 20: Task Information Table - Cash Flow Statement Generation in Accounting

Task Name	{α,β,γ}	Task Description	Input	Output
Find Record Error-Transaction TYPE Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type
Find Record Error-Transaction DATE Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Date; Original Date
Find Record Error-Transaction PAYMENT/RECEIPT_STATUS Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Payment/Receipt Status; Original Payment/Receipt Status
Find Record Error-Transaction PAYMENT_METHOD Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Payment Method; Original Payment Method
Find Record Error-Transaction QUANTITY Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Quantity; Original Quantity
Find Record Error-Transaction UNIT_PRICE Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Unit Price; Original Unit Price
Find Record Error-Transaction RECEIVE_METHOD Record Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Receive Method; Original Receive Method
Find Calculation Error-Transaction AMOUNT Calculation Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Amount; Original Amount
Find Calculation Error-Transaction TAX_AMOUNT Calculation Error	{13,1,3}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax Amount; Original Tax Amount
Find Calculation Error-Transaction PROFIT Calculation Error	{13,1,2}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Profit; Original Profit
Find Transaction Approval Mismatch-Transaction Without PREPARER Error	{13,1,2}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Original Preparer
Find Transaction Approval Mismatch-Transaction Without APPROVER Error	{13,1,4}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Preparer; Original Approver

Table 21: Task Information Table - Single-Error in Auditing

Task Name	{α,β,γ}	Task Description	Input	Output
Find Record Error-Transaction TYPE Record Error & Calculation Error-Transaction TAX_AMOUNT Calculation Error	{13,1,4}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax Amount; Original Tax Amount; Recorded Type; Original Type
Find Record Error-Transaction PAYMENT/RECEIPT_STATUS Record Error & Record Error-Transaction QUANTITY Record Error	{13,1,4}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Payment/Receipt Status; Original Payment/Receipt Status; Recorded Quantity; Original Quantity
Find Record Error-Transaction QUANTITY Record Error & Record Error-Transaction TYPE Record Error	{13,1,4}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Quantity; Original Quantity; Recorded Type; Original Type
Find Record Error-Transaction PAYMENT/RECEIPT_STATUS Record Error & Calculation Error-Transaction AMOUNT Calculation Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Payment/Receipt Status; Original Payment/Receipt Status; Recorded Amount; Original Amount
Find Record Error-Transaction RECEIVE_METHOD Record Error & Record Error-Transaction TYPE Record Error	{13,1,7}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Receive Method; Original Receive Method; Recorded Transaction Type; Original Transaction Type
Find Error-TYPE MISCLASSIFICATION Error & RECORDING DELAY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Date; Original Date
Find Error-TYPE MISCLASSIFICATION Error & PRICE ANOMALY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Price; Original Price
Find Error-TYPE MISCLASSIFICATION Error & AMOUNT DISCREPANCY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Amount; Original Amount
Find Error-RECORDING DELAY Error & PRICE ANOMALY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Date; Original Date; Recorded Price; Original Price
Find Error-RECORDING DELAY Error & AMOUNT DISCREPANCY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Date; Original Date; Recorded Amount; Original Amount
Find Error-PRICE ANOMALY Error & AMOUNT DISCREPANCY Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Price; Original Price; Recorded Amount; Original Amount

Table 22: Task Information Table - Double-Error in Auditing

Task Name	{α,β,γ}	Task Description	Input	Output
Find Record Error-Transaction PAYMENT/RECEIPT_STATUS Record Error & Record Error-Transaction QUANTITY Record Error & Calculation Error-Transaction PROFIT Calculation Error	{13,1,5}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Payment/Receipt Status; Original Payment/Receipt Status; Recorded Quantity; Original Quantity; Recorded Profit; Original Profit
Find Error-TYPE MISCLASSIFICATION Error & RECORDING DELAY Error & PRICE ANOMALY Error	{13,1,7}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Date; Original Date; Recorded Price; Original Price
Find Error-TYPE MISCLASSIFICATION Error & RECORDING DELAY Error & AMOUNT DISCREPANCY Error	{13,1,7}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Date; Original Date; Recorded Amount; Original Amount
Find Error-TYPE MISCLASSIFICATION Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error	{13,1,7}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Type; Original Type; Recorded Price; Original Price; Recorded Amount; Original Amount
Find Error-RECORDING DELAY Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error	{13,1,7}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Date; Original Date; Recorded Price; Original Price; Recorded Amount; Original Amount
Find Error-TAX Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & RECORDING DELAY Error	{13,1,9}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax; Original Tax; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Date; Original Date
Find Error-TAX Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & TYPE MISCLASSIFICATION Error	{13,1,9}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax; Original Tax; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Type; Original Type
Find Error-TAX Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & QUANTITY MISMATCH Error	{13,1,9}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax; Original Tax; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Quantity; Original Quantity
Find Error-PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & RECORDING DELAY Error & QUANTITY MISMATCH Error	{13,1,9}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Date; Original Date; Recorded Quantity; Original Quantity
Find Error-TAX Error & PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & RECORDING DELAY Error & TYPE MISCLASSIFICATION Error	{13,1,11}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax; Original Tax; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Date; Original Date; Recorded Type; Original Type
Find Error-TAX Error & PRICE ANOMALY Error & RECORDING DELAY Error & TYPE MISCLASSIFICATION Error & QUANTITY MISMATCH Error	{13,1,11}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Tax; Original Tax; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Date; Original Date; Recorded Type; Original Type
Find Error-PRICE ANOMALY Error & AMOUNT DISCREPANCY Error & RECORDING DELAY Error & TYPE MISCLASSIFICATION Error & QUANTITY MISMATCH Error	{13,1,11}	Analyze each transaction entry and identify any internal inconsistencies or Errors in the recorded information. Some fields—such as invoice—are generally considered more reliable due to their standardized and regulated nature. For each inconsistency you find, output transaction ID, (the incorrect field(s),) their recorded values, and your best estimate of the correct value(s) based on the other fields in that row.	All transactions data	ID; Recorded Price; Original Price; Recorded Amount; Original Amount; Recorded Date; Original Date; Recorded Type; Original Type; Recorded Quantity; Original Quantity

Table 23: Task Information Table - Multi-Error in Auditing

Task Name	{α,β,γ}	Task Description	Input	Output
Analyze Balance Sheet-Calculate Current Ratio	{2,1,1}	Based on the balance sheet, calculate the Current Ratio as of the end of the reporting period	Balance sheet	The value of Current Ratio
Analyze Balance Sheet-calculate Quick Ratio	{6,1,1}	Based on the balance sheet as of the end of the reporting period	Balance Sheet	The value of Quick Ratio
Analyze Balance Sheet-calculate Debt to Asset Ratio	{2,1,1}	Based on the balance sheet as of the end of the reporting period	Balance Sheet	The value of Debt to Asset Ratio
Analyze Balance Sheet-calculate Debt to Equity Ratio	{2,1,1}	Based on the balance sheet as of the end of the reporting period	Balance Sheet	The value of Debt to Equity Ratio
Analyze Income Statement-Gross Profit Margin	{2,1,1}	Based on the income statement, calculate the Gross Profit Margin	Income Statement	The value of Gross Profit Margin
Analyze Income Statement-Net Profit Margin	{2,1,1}	Based on the income statement, calculate the Net Profit Margin	Income Statement	The value of Net Profit Margin
Analyze Cash Flow Statement-FCF	{2,1,1}	Based on the cash flow statement, calculate the FCF	Cash Flow Statement	The value of FCF
Analyze Cash Flow Statement-Net Cash Ratio	{2,1,1}	Based on the cash flow statement, calculate the Net Cash Ratio	Cash Flow Statement	The value of Net Cash Ratio
Analyze Financial Statement-Cash to Current Debt Ratio	{2,1,1}	Based on the three financial statements, calculate the Cash to Current Debt Ratio	All Financial Statements	The value of Cash to Current Debt Ratio
Analyze Financial Statement-Operating Cash Flow to Current Liabilities Ratio	{3,3,1}	Based on the three financial statements, calculate the Operating Cash Flow to Current Liabilities Ratio	All Financial Statements	The value of Operating Cash Flow to Current Liabilities Ratio
Analyze Financial Statement-ROA	{3,3,1}	Based on the three financial statements, calculate the ROA	All Financial Statements	The value of ROA
Analyze Financial Statement-ROE	{3,3,1}	Based on the three financial statements, calculate the ROE	All Financial Statements	The value of ROE
Analyze Financial Statement-Inventory Turnover Ratio	{3,3,1}	Based on the three financial statements, calculate the Inventory Turnover Ratio	All Financial Statements	The value of Inventory Turnover Ratio
Analyze Financial Statement-Accounts Receivable Turnover Ratio	{3,3,1}	Based on the three financial statements, calculate the Accounts Receivable Turnover Ratio	All Financial Statements	The value of Accounts Receivable Turnover Ratio
Analyze Financial Statement-Current Assets Turnover Ratio	{3,3,1}	Based on the three financial statements, calculate the Current Assets Turnover Ratio	All Financial Statements	The value of Current Assets Turnover Ratio
Analyze Financial Statement-Total Asset Turnover Ratio	{3,3,1}	Based on the three financial statements, calculate the Total Asset Turnover Ratio	All Financial Statements	The value of Total Asset Turnover Ratio
Analyze Financial Statement-Cash Flow to Debt Ratio	{2,3,1}	Based on the three financial statements, calculate the Cash Flow to Debt Ratio	All Financial Statements	The value of Cash Flow to Debt Ratio
Analyze Financial Statement-Operating Cash Flow Ratio	{2,3,1}	Based on the three financial statements, calculate the Operating Cash Flow Ratio	All Financial Statements	The value of Operating Cash Flow Ratio

Table 24: Task Information Table - Single-Capability in Consulting

Analyze Financial Statement-Current Ratio & Inventory Turnover Ratio	{5,3,2}	Based on the three financial statements, calculate the Current Ratio and Inventory Turnover Ratio	All Financial Statements	The value of Current Ratio and Inventory Turnover Ratio
Analyze Financial Statement-Gross Profit Margin & Operating Cash Flow Ratio	{4,3,2}	Based on the three financial statements, calculate the Gross Profit Margin and Operating Cash Flow Ratio	All Financial Statements	The value of Gross Profit Margin and Operating Cash Flow Ratio
Analyze Financial Statement-FCF & Current Assets Turnover Ratio	{5,3,2}	Based on the three financial statements, calculate the FCF and Current Assets Turnover Ratio	All Financial Statements	The value of FCF and Current Assets Turnover Ratio
Analyze Financial Statement-Quick Ratio & Net Profit Margin	{8,3,2}	Based on the three financial statements, calculate the Quick Ratio and Net Profit Margin	All Financial Statements	The value of Quick Ratio and Net Profit Margin
Analyze Financial Statement-Gross Profit Margin & Current Liabilities Ratio	{4,3,2}	Based on the three financial statements, calculate the Gross Profit Margin and Current Liabilities Ratio	All Financial Statements	The value of Gross Profit Margin and Current Liabilities Ratio
Analyze Financial Statement-Debt to Asset Ratio & Net Cash Ratio	{4,3,2}	Based on the three financial statements, calculate the Debt to Asset Ratio and Net Cash Ratio	All Financial Statements	The value of Debt to Asset Ratio and Net Cash Ratio
Analyze Financial Statement-Debt to Equity Ratio & Net Profit Margin & Operating Cash Flow to Current Liabilities Ratio	{6,3,3}	Based on the three financial statements, calculate the Debt to Equity Ratio, Net Profit Margin and Operating Cash Flow to Current Liabilities Ratio	All Financial Statements	The value of Debt to Equity Ratio, Net Profit Margin and Operating Cash Flow to Current Liabilities Ratio
Analyze Financial Statement-ROE & Debt to Asset Ratio & Gross Profit Margin	{7,3,3}	Based on the three financial statements, calculate the ROE, Debt to Asset Ratio and Gross Profit Margin	All Financial Statements	The value of ROE, Debt to Asset Ratio and Gross Profit Margin
Analyze Financial Statement-Net Cash Ratio & Turnover Ratio & Quick Ratio	{11,3,3}	Based on the three financial statements, calculate the Net Cash Ratio, Turnover Ratio and Quick Ratio	All Financial Statements	The value of Net Cash Ratio, Turnover Ratio and Quick Ratio
Analyze Financial Statement-Debt to Asset Ratio & Gross Profit Margin & Operating Cash Flow Ratio	{6,3,3}	Based on the three financial statements, calculate the Debt to Asset Ratio, Gross Profit Margin and Operating Cash Flow Ratio	All Financial Statements	The value of Debt to Asset Ratio, Gross Profit Margin and Operating Cash Flow Ratio
Analyze Financial Statement-Debt to Equity Ratio & Net Profit Margin & ROA & Accounts Receivable Turnover Ratio	{10,3,4}	Based on the three financial statements, calculate the Debt to Equity Ratio, Net Profit Margin, ROA and Accounts Receivable Turnover Ratio	All Financial Statements	The value of Debt to Equity Ratio, Net Profit Margin, ROA and Accounts Receivable Turnover Ratio
Analyze Financial Statement-Current Ratio & Quick Ratio & Debt to Asset Ratio & Debt to Equity Ratio & Cash Flow to Debt Ratio	{14,3,5}	Based on the three financial statements, calculate the Current Ratio, Quick Ratio, Debt to Asset Ratio, Debt to Equity Ratio, Cash Flow to Debt Ratio	All Financial Statements	The value of Current Ratio, Quick Ratio, Debt to Asset Ratio, Debt to Equity Ratio, Cash Flow to Debt Ratio
Analyze Financial Statement-Accounts Receivable Turnover Ratio & Operating Cash Flow to Current Liabilities Ratio & Operating Cash Flow Ratio & Total Asset Turnover Ratio & Debt to Equity Ratio	{12,3,5}	Based on the three financial statements, calculate the Accounts Receivable Turnover Ratio, Operating Cash Flow to Current Liabilities Ratio, Operating Cash Flow Ratio, Total Asset Turnover Ratio and Debt to Equity Ratio	All Financial Statements	The value of Accounts Receivable Turnover Ratio, Operating Cash Flow to Current Liabilities Ratio, Operating Cash Flow Ratio, Total Asset Turnover Ratio and Debt to Equity Ratio
Analyze Financial Statement-FCF & ROA & ROE & Net Cash Ratio & Net Profit Margin & Gross Profit Margin	{14,3,6}	Based on the three financial statements, calculate the FCF, ROA, ROE, Net Cash Ratio, Net Profit Margin and Gross Profit Margin	All Financial Statements	The value of FCF, ROA, ROE, Net Cash Ratio, Net Profit Margin and Gross Profit Margin
Analyze Financial Statement-Operating Cash Flow Ratio & Cash Flow to Debt Ratio & Inventory Turnover Ratio & Debt to Equity Ratio & Quick Ratio & Current Ratio	{17,3,6}	Based on the three financial statements, calculate the Operating Cash Flow Ratio, Cash Flow to Debt Ratio, Inventory Turnover Ratio, Debt to Equity Ratio, Quick Ratio and Current Ratio	All Financial Statements	The value of Operating Cash Flow Ratio, Cash Flow to Debt Ratio, Inventory Turnover Ratio, Debt to Equity Ratio, Quick Ratio and Current Ratio
Analyze Financial Statement-Operating Cash Flow to Current Liabilities Ratio & Debt to Equity Ratio & Total Asset Turnover Ratio & Quick Ratio & Operating Cash Flow Ratio & ROE & Accounts Receivable Turnover Ratio	{21,3,7}	Based on the three financial statements, calculate the Operating Cash Flow to Current Liabilities Ratio, Debt to Equity Ratio, Total Asset Turnover Ratio, Quick Ratio, Operating Cash Flow Ratio, ROE and Accounts Receivable Turnover Ratio	All Financial Statements	The value of Operating Cash Flow to Current Liabilities Ratio, Debt to Equity Ratio, Total Asset Turnover Ratio, Quick Ratio, Operating Cash Flow Ratio, ROE and Accounts Receivable Turnover Ratio
Analyze Financial Statement-Current Ratio & Gross Profit Margin & Debt to Asset Ratio & Net Profit Margin & Cash to Current Debt Ratio & FCF & ROA	{15,3,7}	Based on the three financial statements, calculate the Current Ratio, Gross Profit Margin, Debt to Asset Ratio, Net Profit Margin, Cash to Current Debt Ratio, FCF and ROA	All Financial Statements	The value of Current Ratio, Gross Profit Margin, Debt to Asset Ratio, Net Profit Margin, Cash to Current Debt Ratio, FCF and ROA

Table 25: Task Information Table - Multi-Capability in Consulting

F Extended Experiment Result

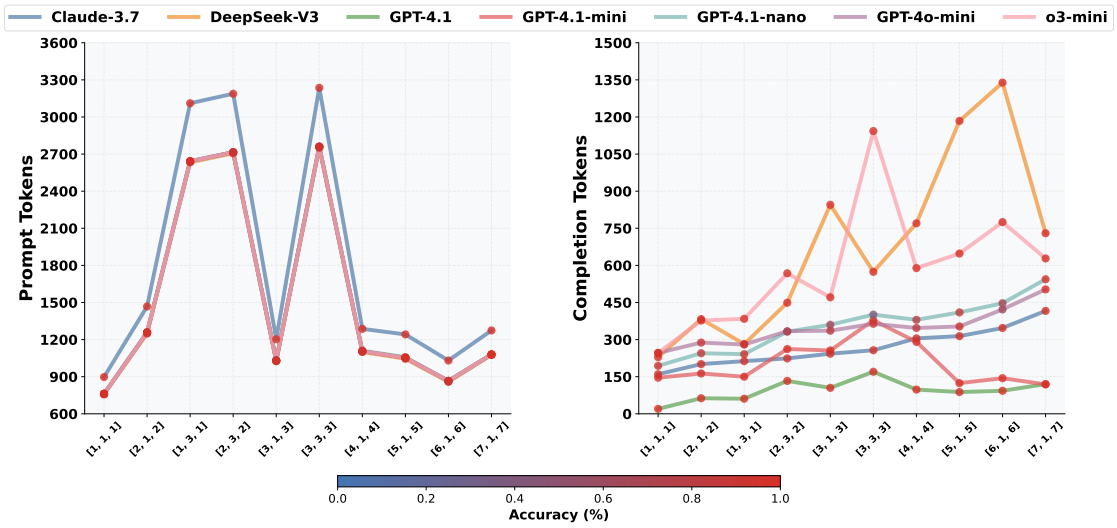


Figure 7: Financial Literacy prompt and completion Token Result

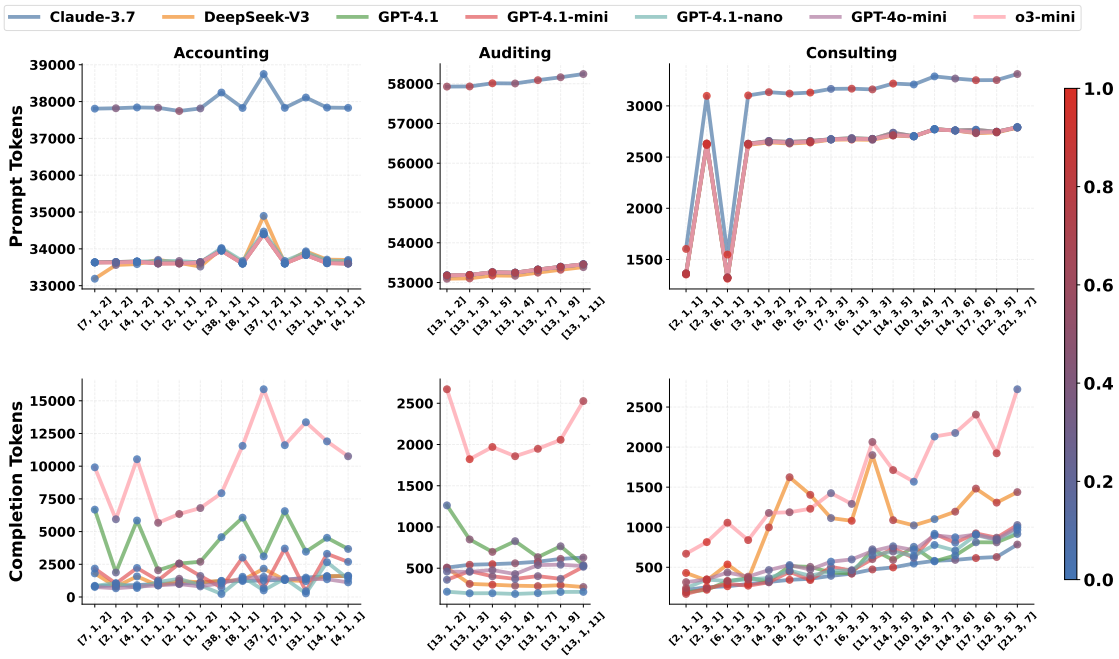


Figure 8: Main task prompt and completion result for model Comparison

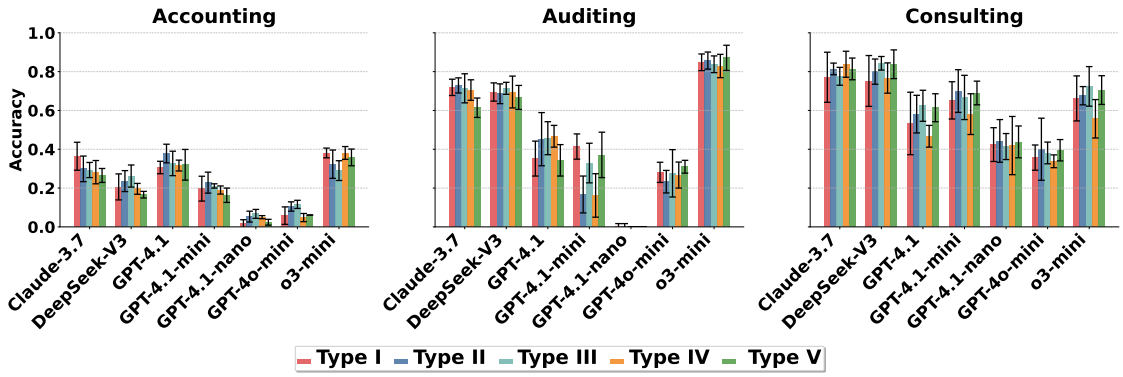


Figure 9: Performance comparison of LLMs across different company configuration

Metrics	DeepSeek-V3	GPT-4.1	GPT-4o-mini
Transaction-400 Long Cycle			
Financial Literacy	99.22% ±0.89%	99.53% ±0.73%	88.91% ±1.94%
Accounting	15.04% ±3.49%	20.27% ±4.18%	3.81% ±1.76%
Auditing	62.35% ±6.54%	37.05% ±11.56%	6.19% ±3.68%
Consulting	78.10% ±8.09%	61.52% ±7.97%	39.52% ±8.19%
Transaction-200 Short Cycle			
Financial Literacy	99.06% ±0.88%	99.90% ±0.40%	89.01% ±2.19%
Accounting	21.33% ±5.63%	32.93% ±5.59%	7.76% ±3.73%
Auditing	69.14% ±5.53%	41.43% ±10.31%	27.33% ±7.22%
Consulting	80.00% ±8.52%	56.38% ±11.00%	37.43% ±8.33%

Table 26: Model accuracy for different operation time

Model / Company Type	Financial Literacy	Accounting	Auditing	Consulting
Claude-3.7-Sonnet				
Type I	99.38%	35.10%	73.14%	74.86%
Type II	99.74%	29.66%	72.86%	81.43%
Type III	99.48%	29.35%	71.43%	77.62%
Type IV	100.00%	26.61%	70.48%	83.81%
Type V	99.33%	28.53%	62.04%	82.45%
DeepSeek-V3				
Type I	98.75%	21.80%	68.57%	73.71%
Type II	99.22%	23.57%	68.57%	80.00%
Type III	99.48%	25.20%	71.43%	84.29%
Type IV	98.96%	19.31%	69.52%	76.67%
Type V	98.88%	17.27%	67.76%	83.67%
GPT-4.1				
Type I	99.69%	29.80%	36.00%	56.00%
Type II	99.74%	37.76%	45.24%	58.10%
Type III	100.00%	32.65%	45.71%	62.38%
Type IV	100.00%	31.63%	46.67%	46.67%
Type V	100.00%	32.36%	33.88%	58.37%
GPT-4.1-mini				
Type I	98.13%	21.22%	42.86%	63.43%
Type II	98.96%	22.79%	16.67%	70.00%
Type III	97.92%	21.09%	32.86%	66.67%
Type IV	97.92%	19.05%	16.19%	58.10%
Type V	97.77%	15.74%	36.73%	69.80%
GPT-4.1-nano				
Type I	86.25%	0.93%	0.57%	40.00%
Type II	83.33%	5.30%	0.48%	44.29%
Type III	83.59%	6.41%	0.00%	41.43%
Type IV	88.28%	4.90%	0.00%	38.62%
Type V	85.04%	2.56%	0.00%	45.31%
GPT-4o-mini				
Type I	89.38%	5.31%	29.14%	36.57%
Type II	89.06%	10.54%	23.33%	40.00%
Type III	89.06%	11.56%	27.62%	38.10%
Type IV	88.02%	4.76%	26.67%	33.81%
Type V	89.51%	6.41%	29.80%	38.37%
o3-mini				
Type I	100.00%	37.55%	86.29%	64.00%
Type II	100.00%	32.31%	85.71%	67.62%
Type III	100.00%	29.01%	83.81%	72.38%
Type IV	100.00%	38.10%	82.86%	55.71%
Type V	100.00%	36.55%	85.71%	71.43%

Table 27: Model Comparison across Different Company Types

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	100%	160	897
	[1,3,1]	100%	213	3,111
	[2,1,2]	99%	201	1,468
	[2,3,2]	100%	224	3,188
	[3,1,3]	100%	243	1,203
	[3,3,3]	100%	257	3,236
	[4,1,4]	97.78%	305	1,287
	[5,1,5]	100%	314	1,242
	[6,1,6]	100%	347	1,031
	[7,1,7]	100%	416	1,274
Accounting	[1,1,1]	41.9%	1,015	37,832
	[1,1,2]	30.45%	1,102	37,814
	[14,1,1]	0%	1,588	37,837
	[2,1,1]	52%	1,018	37,744
	[2,1,2]	39.33%	892	37,819
	[31,1,1]	0%	1,483	38,110
	[37,1,2]	0%	1,262	38,748
	[38,1,1]	0%	1,213	38,246
	[4,1,1]	0%	1,595	37,830
	[4,1,2]	0%	906	37,841
	[7,1,1]	0%	1,310	37,830
	[7,1,2]	0%	822	37,808
	[8,1,1]	0%	1,220	37,827
Auditing	[13,1,11]	42.22%	629	58,243
	[13,1,2]	61.67%	508	57,927
	[13,1,3]	76.33%	542	57,931
	[13,1,4]	52.5%	560	58,006
	[13,1,5]	89.05%	549	58,011
	[13,1,7]	76%	579	58,089
	[13,1,9]	52.5%	613	58,162
Consulting	[10,3,4]	0%	546	3,209
	[11,3,3]	76.67%	473	3,161
	[12,3,5]	66.67%	628	3,252
	[14,3,5]	93.33%	498	3,218
	[14,3,6]	36.67%	592	3,267
	[15,3,7]	0%	580	3,288
	[17,3,6]	83.33%	615	3,251
	[2,1,1]	91.11%	218	1,603
	[2,3,1]	98.33%	250	3,098
	[21,3,7]	46.67%	784	3,311
	[3,3,1]	95%	291	3,102
	[4,3,2]	83.33%	317	3,134
	[5,3,2]	98.33%	356	3,130
	[6,1,1]	100%	263	1,548
	[6,3,3]	86.67%	420	3,168
	[7,3,3]	10%	393	3,166
	[8,3,2]	90%	343	3,120

Table 28: Claude-3.7-Sonnet Model Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	100%	229	757
	[1,3,1]	100%	282	2,633
	[2,1,2]	97%	383	1,249
	[2,3,2]	100%	449	2,706
	[3,1,3]	100%	845	1,025
	[3,3,3]	100%	574	2,751
	[4,1,4]	100%	770	1,100
	[5,1,5]	100%	1,184	1,045
	[6,1,6]	100%	1,339	858
	[7,1,7]	100%	730	1,076
Accounting	[1,1,1]	23.39%	878	33,701
	[1,1,2]	25.87%	1,123	33,519
	[14,1,1]	0%	1,586	33,708
	[2,1,1]	52.94%	1,221	33,618
	[2,1,2]	30.87%	792	33,562
	[31,1,1]	0%	1,214	33,933
	[37,1,2]	0%	2,164	34,894
	[38,1,1]	0%	1,237	34,003
	[4,1,1]	1.85%	1,625	33,699
	[4,1,2]	0%	1,572	33,586
	[7,1,1]	0%	1,292	33,604
	[7,1,2]	0%	1,805	33,189
	[8,1,1]	0%	1,285	33,615
Auditing	[13,1,11]	45.56%	272	53,387
	[13,1,2]	50%	503	53,096
	[13,1,3]	72.33%	310	53,105
	[13,1,4]	49.17%	288	53,170
	[13,1,5]	87.62%	300	53,177
	[13,1,7]	70.67%	283	53,249
	[13,1,9]	74.17%	294	53,321
Consulting	[10,3,4]	16.67%	1,023	2,701
	[11,3,3]	43.33%	1,900	2,668
	[12,3,5]	80%	1,307	2,740
	[14,3,5]	73.33%	1,088	2,710
	[14,3,6]	66.67%	1,194	2,760
	[15,3,7]	0%	1,101	2,771
	[17,3,6]	70%	1,482	2,733
	[2,1,1]	92.59%	429	1,353
	[2,3,1]	96.67%	335	2,617
	[21,3,7]	56.67%	1,438	2,787
	[3,3,1]	96.11%	350	2,619
	[4,3,2]	81.11%	998	2,643
	[5,3,2]	83.33%	1,404	2,642
	[6,1,1]	100%	535	1,312
	[6,3,3]	91.67%	1,080	2,671
	[7,3,3]	23.33%	1,114	2,670
	[8,3,2]	73.33%	1,624	2,632

Table 29: DeepSeek-V3 Model Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	100%	20	762
	[1,3,1]	100%	61	2,642
	[2,1,2]	99.67%	63	1,257
	[2,3,2]	100%	133	2,715
	[3,1,3]	100%	105	1,030
	[3,3,3]	100%	170	2,760
	[4,1,4]	100%	98	1,106
	[5,1,5]	100%	88	1,053
	[6,1,6]	100%	93	866
	[7,1,7]	100%	120	1,080
Accounting	[1,1,1]	43.51%	2,044	33,607
	[1,1,2]	45%	2,694	33,638
	[14,1,1]	0%	4,525	33,618
	[2,1,1]	53.33%	2,556	33,613
	[2,1,2]	25.33%	1,872	33,642
	[31,1,1]	0%	3,464	33,837
	[37,1,2]	0%	3,106	34,403
	[38,1,1]	0%	4,575	33,957
	[4,1,1]	6.67%	3,674	33,604
	[4,1,2]	0%	5,847	33,659
	[7,1,1]	0%	6,565	33,606
	[7,1,2]	0%	6,678	33,635
	[8,1,1]	0%	6,065	33,602
Auditing	[13,1,11]	34.44%	538	53,459
	[13,1,2]	13.33%	1,262	53,180
	[13,1,3]	45.67%	849	53,191
	[13,1,4]	30%	828	53,248
	[13,1,5]	53.81%	698	53,259
	[13,1,7]	46.67%	634	53,326
	[13,1,9]	33.33%	766	53,396
Consulting	[10,3,4]	6.67%	755	2,704
	[11,3,3]	23.33%	724	2,677
	[12,3,5]	36.67%	811	2,747
	[14,3,5]	56.67%	592	2,737
	[14,3,6]	0%	647	2,760
	[15,3,7]	0%	574	2,772
	[17,3,6]	13.33%	812	2,766
	[2,1,1]	82.96%	188	1,364
	[2,3,1]	95%	219	2,632
	[21,3,7]	6.67%	916	2,791
	[3,3,1]	75%	366	2,630
	[4,3,2]	37.78%	327	2,657
	[5,3,2]	46.67%	505	2,658
	[6,1,1]	90%	325	1,325
	[6,3,3]	46.67%	424	2,685
	[7,3,3]	10%	435	2,674
	[8,3,2]	43.33%	514	2,648

Table 30: GPT-4.1 Model Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	97.96%	146	762
	[1,3,1]	100%	150	2,642
	[2,1,2]	95.83%	163	1,257
	[2,3,2]	100%	262	2,715
	[3,1,3]	100%	256	1,030
	[3,3,3]	100%	378	2,760
	[4,1,4]	100%	291	1,106
	[5,1,5]	100%	124	1,053
	[6,1,6]	100%	144	866
	[7,1,7]	100%	119	1,080
Accounting	[1,1,1]	24.91%	1,271	33,603
	[1,1,2]	25.28%	1,608	33,634
	[14,1,1]	0%	3,298	33,611
	[2,1,1]	40%	2,555	33,602
	[2,1,2]	22%	1,030	33,638
	[31,1,1]	0%	439	33,834
	[37,1,2]	0%	681	34,400
	[38,1,1]	0%	715	33,952
	[4,1,1]	1.67%	2,685	33,601
	[4,1,2]	0%	2,219	33,651
	[7,1,1]	0%	3,705	33,602
	[7,1,2]	0%	2,172	33,632
	[8,1,1]	0%	3,017	33,599
Auditing	[13,1,11]	22.22%	517	53,459
	[13,1,2]	16.67%	360	53,180
	[13,1,3]	29.67%	465	53,191
	[13,1,4]	19.17%	371	53,248
	[13,1,5]	38.10%	402	53,259
	[13,1,7]	33.33%	404	53,326
	[13,1,9]	25.83%	371	53,396
Consulting	[10,3,4]	23.33%	617	2,704
	[11,3,3]	56.67%	602	2,677
	[12,3,5]	46.67%	869	2,747
	[14,3,5]	76.67%	735	2,715
	[14,3,6]	3.33%	807	2,760
	[15,3,7]	0%	918	2,772
	[17,3,6]	63.33%	923	2,740
	[2,1,1]	80%	167	1,360
	[2,3,1]	66.67%	234	2,626
	[21,3,7]	23.33%	1,027	2,791
	[3,3,1]	91.67%	270	2,627
	[4,3,2]	41.11%	312	2,651
	[5,3,2]	85%	340	2,650
	[6,1,1]	93.33%	293	1,321
	[6,3,3]	63.33%	458	2,677
	[7,3,3]	23.33%	505	2,674
	[8,3,2]	70%	444	2,641

Table 31: GPT-4.1-mini Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	100%	246	761
	[1,3,1]	100%	384	2,641
	[2,1,2]	100%	377	1,256
	[2,3,2]	100%	568	2,714
	[3,1,3]	100%	471	1,029
	[3,3,3]	100%	1,143	2,759
	[4,1,4]	100%	589	1,105
	[5,1,5]	100%	648	1,052
	[6,1,6]	100%	775	865
	[7,1,7]	100%	628	1,079
Accounting	[1,1,1]	47.02%	5,673	33,602
	[1,1,2]	42.62%	6,798	33,628
	[14,1,1]	3.33%	11,896	33,610
	[2,1,1]	55%	6,345	33,601
	[2,1,2]	24.83%	5,943	33,621
	[31,1,1]	0%	13,361	33,833
	[37,1,2]	0%	15,881	34,399
	[38,1,1]	0%	7,936	33,951
	[4,1,1]	28.33%	10,756	33,600
	[4,1,2]	0%	10,533	33,650
	[7,1,1]	0%	11,607	33,601
	[7,1,2]	0%	9,914	33,631
	[8,1,1]	0%	11,553	33,598
Auditing	[13,1,11]	68.89%	2,527	53,458
	[13,1,2]	81.67%	2,670	53,179
	[13,1,3]	93%	1,822	53,190
	[13,1,4]	78.33%	1,858	53,247
	[13,1,5]	87.62%	1,970	53,258
	[13,1,7]	81.33%	1,948	53,325
	[13,1,9]	84.17%	2,057	53,395
Consulting	[10,3,4]	0%	1,568	2,703
	[11,3,3]	53.33%	2,064	2,676
	[12,3,5]	80%	1,923	2,746
	[14,3,5]	76.67%	1,713	2,714
	[14,3,6]	13.33%	2,176	2,759
	[15,3,7]	0%	2,132	2,771
	[17,3,6]	66.67%	2,406	2,739
	[2,1,1]	84.44%	669	1,359
	[2,3,1]	95%	814	2,625
	[21,3,7]	10%	2,722	2,790
	[3,3,1]	86.11%	839	2,626
	[4,3,2]	50%	1,178	2,650
	[5,3,2]	86.67%	1,230	2,649
	[6,1,1]	86.67%	1,056	1,320
	[6,3,3]	36.67%	1,291	2,676
	[7,3,3]	23.33%	1,425	2,673
	[8,3,2]	53.33%	1,187	2,640

Table 32: o3-mini Model Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	96.48%	246	762
	[1,3,1]	87.50%	280	2,642
	[2,1,2]	89.67%	288	1,257
	[2,3,2]	67.78%	334	2,715
	[3,1,3]	80.00%	336	1,030
	[3,3,3]	91.11%	364	2,760
	[4,1,4]	78.89%	347	1,106
	[5,1,5]	91.11%	353	1,053
	[6,1,6]	60.00%	422	866
	[7,1,7]	100%	503	1,080
Accounting	[1,1,1]	7.37%	890	33,607
	[1,1,2]	14.17%	808	33,638
	[14,1,1]	0%	1,357	33,618
	[2,1,1]	26.67%	984	33,613
	[2,1,2]	3.33%	665	33,642
	[31,1,1]	0%	1,384	33,837
	[37,1,2]	0%	1,537	34,403
	[38,1,1]	0%	1,053	33,957
	[4,1,1]	0%	1,119	33,604
	[4,1,2]	0%	788	33,659
	[7,1,1]	0%	1,346	33,606
	[7,1,2]	0%	767	33,635
	[8,1,1]	0%	1,428	33,602
Auditing	[13,1,11]	14.44%	526	53,459
	[13,1,2]	20.00%	460	53,180
	[13,1,3]	24.00%	456	53,191
	[13,1,4]	33.33%	429	53,248
	[13,1,5]	36.67%	478	53,259
	[13,1,7]	31.33%	538	53,326
	[13,1,9]	21.67%	543	53,396
Consulting	[10,3,4]	0%	714	2,704
	[11,3,3]	16.67%	702	2,677
	[12,3,5]	23.33%	849	2,747
	[14,3,5]	13.33%	762	2,738
	[14,3,6]	0%	874	2,760
	[15,3,7]	0%	898	2,772
	[17,3,6]	0%	909	2,767
	[2,1,1]	58.89%	314	1,364
	[2,3,1]	85.00%	349	2,632
	[21,3,7]	0%	999	2,791
	[3,3,1]	57.22%	383	2,631
	[4,3,2]	21.11%	465	2,658
	[5,3,2]	36.67%	472	2,658
	[6,1,1]	40.00%	434	1,325
	[6,3,3]	15.00%	598	2,685
	[7,3,3]	0%	570	2,674
	[8,3,2]	6.67%	528	2,649

Table 33: GPT-4o-mini Model Accuracy, Complete Completion and Prompt Token Result

Task Type	{α,β,γ}	Percentage (%)	Completion Token	Prompt Token
Financial Literacy	[1,1,1]	100%	246	761
	[1,3,1]	100%	384	2,641
	[2,1,2]	100%	377	1,256
	[2,3,2]	100%	568	2,714
	[3,1,3]	100%	471	1,029
	[3,3,3]	100%	1,143	2,759
	[4,1,4]	100%	589	1,105
	[5,1,5]	100%	648	1,052
	[6,1,6]	100%	775	865
	[7,1,7]	100%	628	1,079
Accounting	[1,1,1]	47.02%	5,673	33,602
	[1,1,2]	42.62%	6,798	33,628
	[14,1,1]	3.33%	11,896	33,610
	[2,1,1]	55%	6,345	33,601
	[2,1,2]	24.83%	5,943	33,621
	[31,1,1]	0%	13,361	33,833
	[37,1,2]	0%	15,881	34,399
	[38,1,1]	0%	7,936	33,951
	[4,1,1]	28.33%	10,756	33,600
	[4,1,2]	0%	10,533	33,650
	[7,1,1]	0%	11,607	33,601
	[7,1,2]	0%	9,914	33,631
	[8,1,1]	0%	11,553	33,598
Auditing	[13,1,11]	68.89%	2,527	53,458
	[13,1,2]	81.67%	2,670	53,179
	[13,1,3]	93%	1,822	53,190
	[13,1,4]	78.33%	1,858	53,247
	[13,1,5]	87.62%	1,970	53,258
	[13,1,7]	81.33%	1,948	53,325
	[13,1,9]	84.17%	2,057	53,395
Consulting	[10,3,4]	0%	1,568	2,703
	[11,3,3]	53.33%	2,064	2,676
	[12,3,5]	80%	1,923	2,746
	[14,3,5]	76.67%	1,713	2,714
	[14,3,6]	13.33%	2,176	2,759
	[15,3,7]	0%	2,132	2,771
	[17,3,6]	66.67%	2,406	2,739
	[2,1,1]	84.44%	669	1,359
	[2,3,1]	95%	814	2,625
	[21,3,7]	10%	2,722	2,790
	[3,3,1]	86.11%	839	2,626
	[4,3,2]	50%	1,178	2,650
	[5,3,2]	86.67%	1,230	2,649
	[6,1,1]	86.67%	1,056	1,320
	[6,3,3]	36.67%	1,291	2,676
	[7,3,3]	23.33%	1,425	2,673
	[8,3,2]	53.33%	1,187	2,640

Table 34: o3-mini Model Accuracy, Complete Completion and Prompt Token Result

G FinEval

G.1 Models

Type	Models	Version	Provider
Online	GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI
	GPT-4.1	gpt-4.1-2025-04-14	OpenAI
	GPT-4.1-mini	gpt-4.1-mini-2025-04-14	OpenAI
	GPT-4.1-nano	gpt-4.1-nano-2025-04-14	OpenAI
	o3-mini	o3-mini-2025-01-31	OpenAI
	DeepSeek-V3	DeepSeek-V3-250324	Huoshan
	Claude-3.7-Sonnet	claude-3-7-sonnet-20250219	Anthropic

Table 35: API-based LLMs considered in this paper via *FinEval*.

G.2 Prompt Template

```
finmaster_template = """
# <task_name> Task Description:
<task_description>

# Examples:
<in_context_examples>
# Problem to Solve:
{"problem": <task_to_solve>}

# Instruction:
Now please solve the above task. Reason step by step and present your answer in the "
    solution" field in the following json format:
```json
{"solution": "___" }
```
"""

example_and_solution = """{"problem": <example_problem>}
{"solution": <example_solution>}
"""
```

H Accuracy of Each Company Type for Specific Task

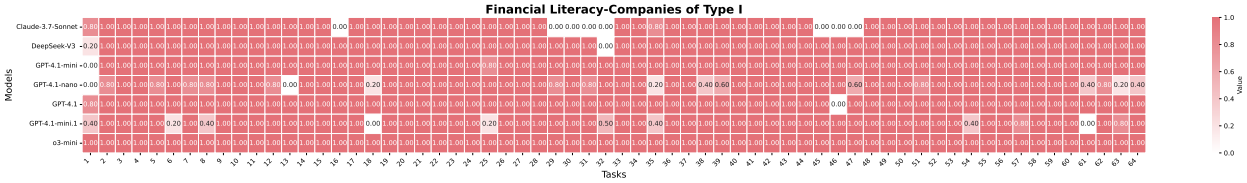


Figure 10: Accuracy of type I companies in financial literacy

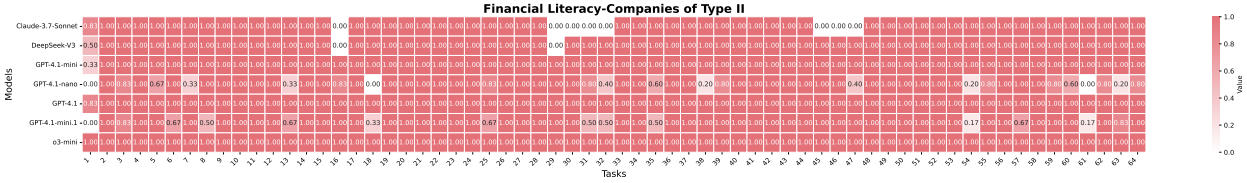


Figure 11: Accuracy of Type II Companies in Financial Literacy

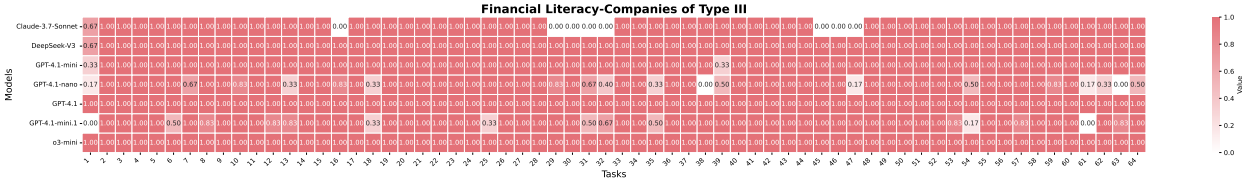


Figure 12: Accuracy of Type III Companies in Financial Literacy

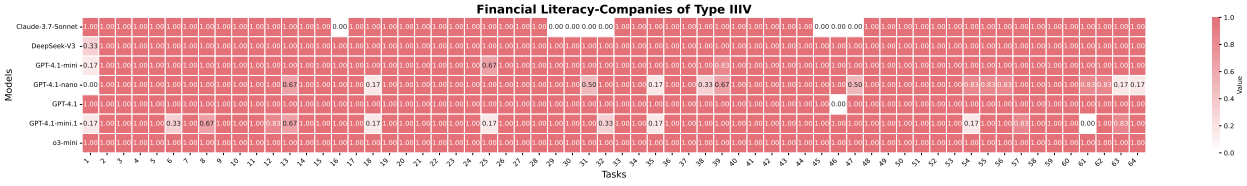


Figure 13: Accuracy of Type IIIV Companies in Financial Literacy

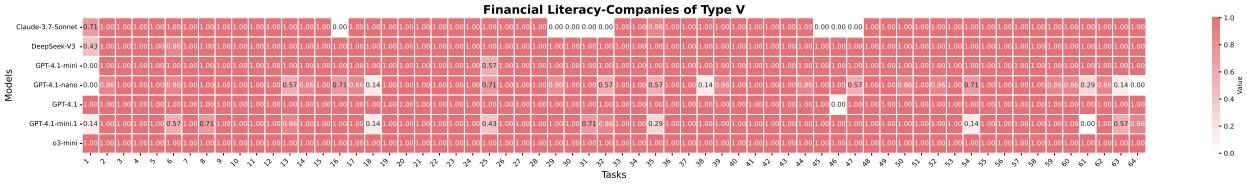


Figure 14: Accuracy of Type V Companies in Financial Literacy

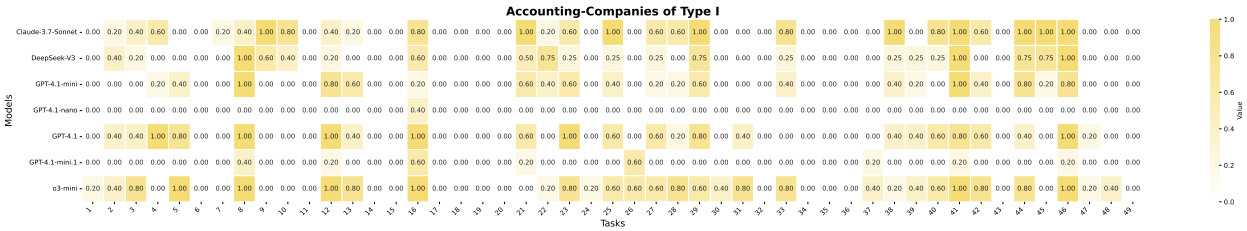


Figure 15: Accuracy of Type I Companies in Accounting

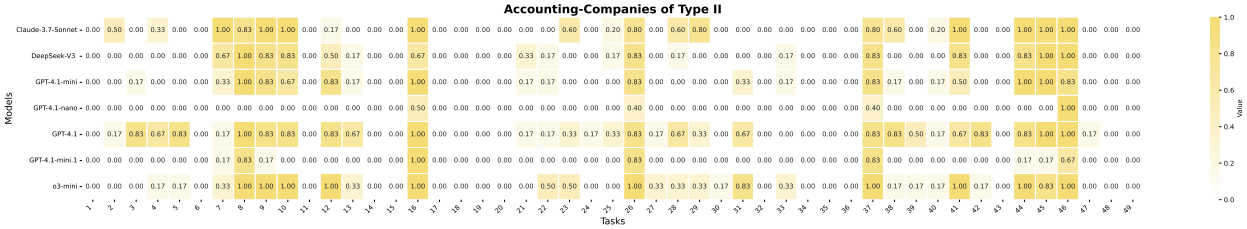


Figure 16: Accuracy of Type II Companies in Accounting

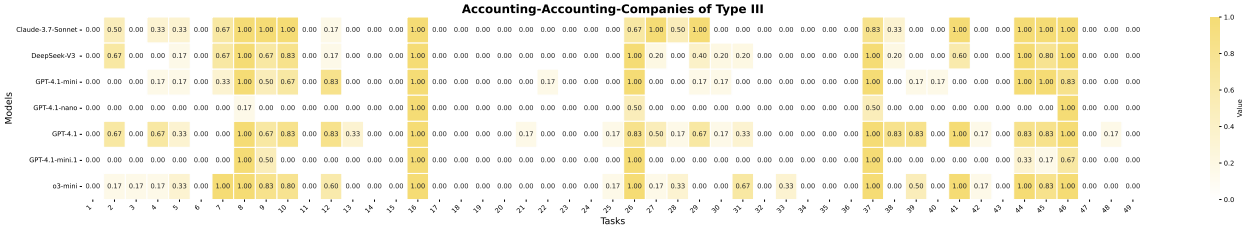


Figure 17: Accuracy of Type III Companies in Accounting

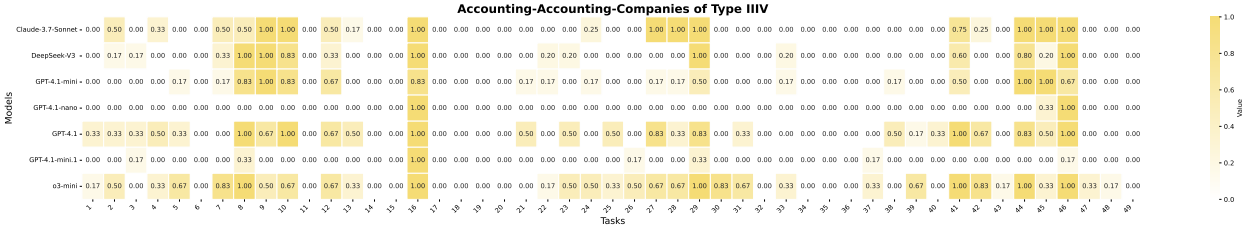


Figure 18: Accuracy of Type IIIV Companies in Accounting

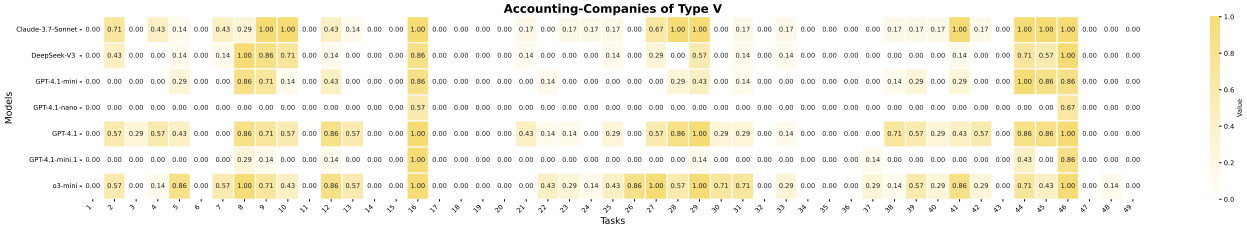


Figure 19: Accuracy of Type V Companies in Accounting

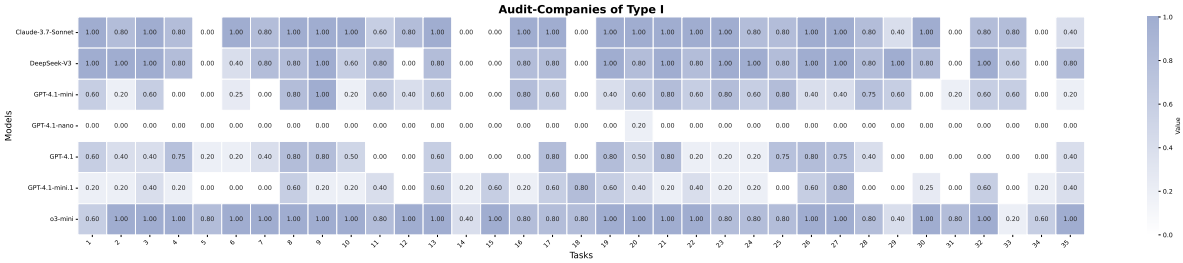


Figure 20: Accuracy of Type I Companies in Audit

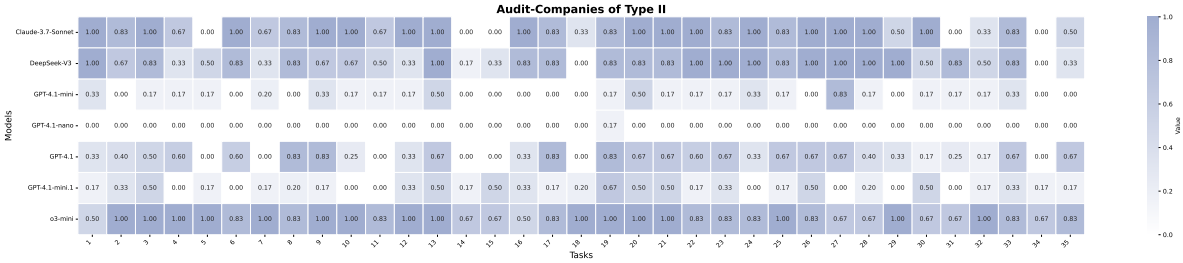


Figure 21: Accuracy of Type II Companies in Audit

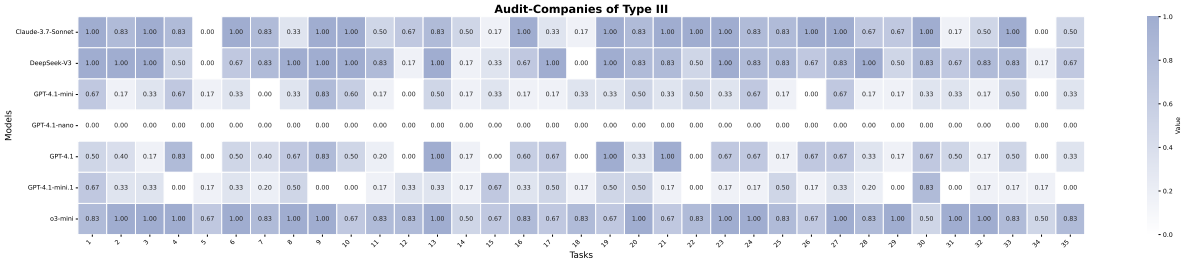


Figure 22: Accuracy of Type III Companies in Audit

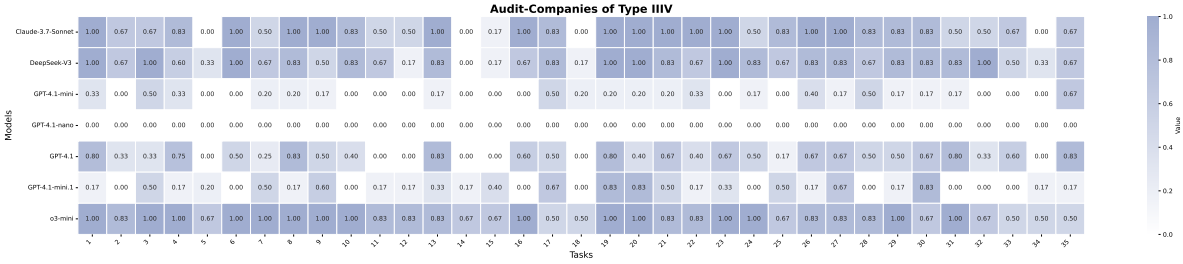


Figure 23: Accuracy of Type IIIV Companies in Audit

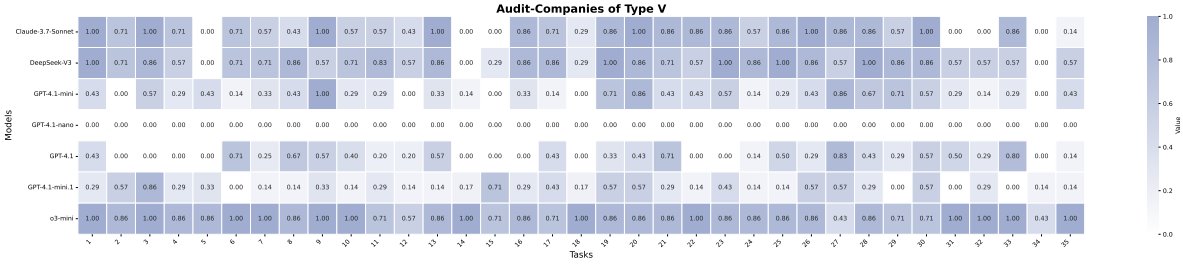


Figure 24: Accuracy of Type V Companies in Audit

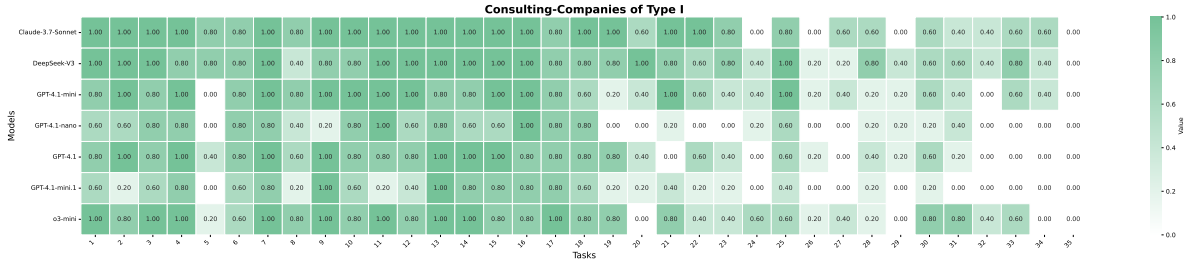


Figure 25: Accuracy of Type I Companies in Consulting

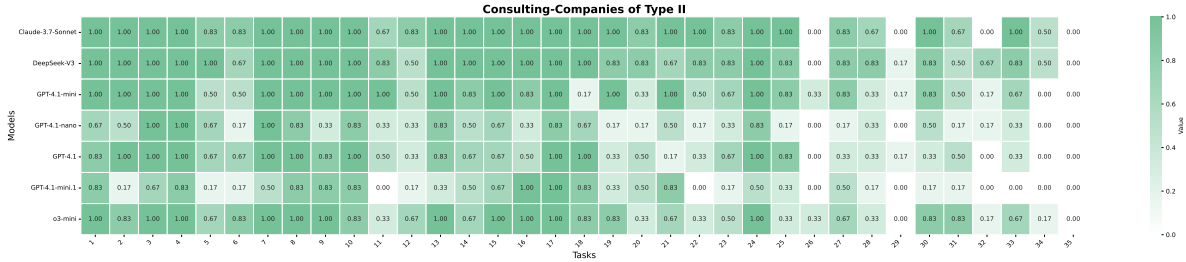


Figure 26: Accuracy of Type II Companies in Consulting

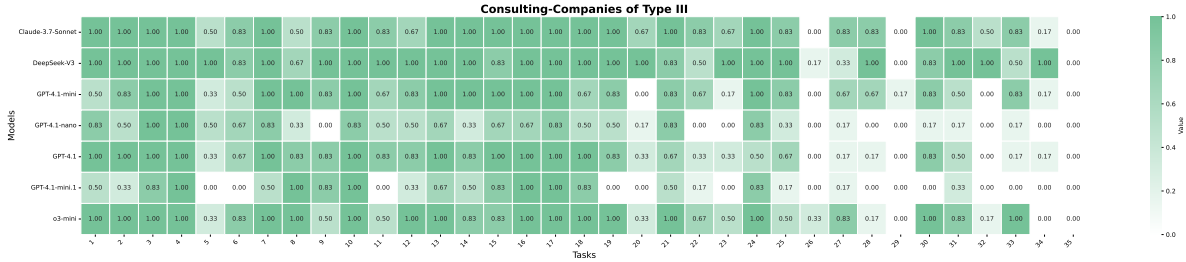


Figure 27: Accuracy of Type III Companies in Consulting

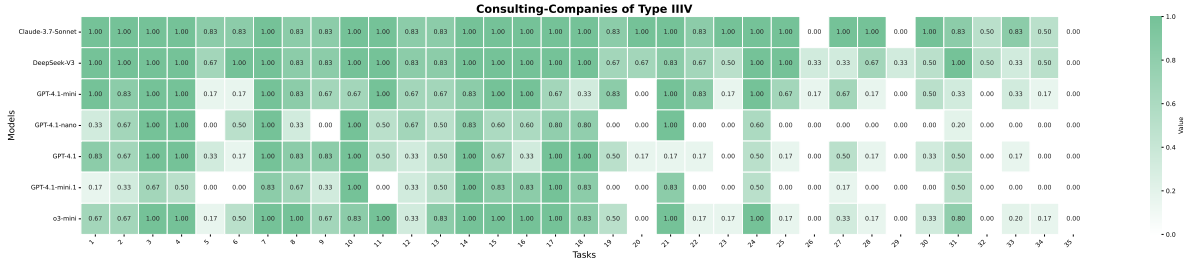


Figure 28: Accuracy of Type IIIV Companies in Consulting

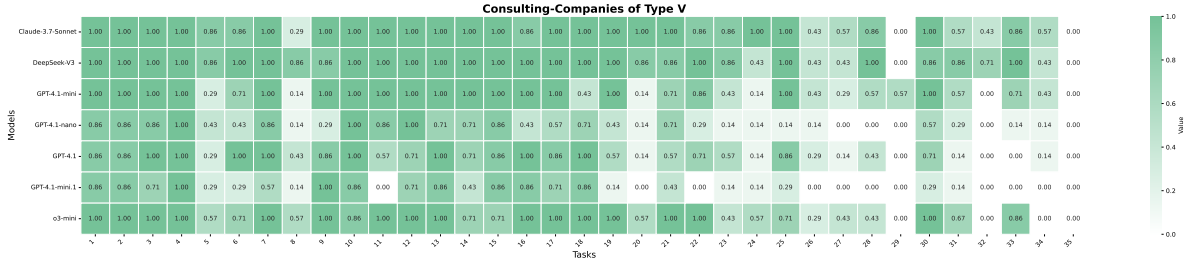


Figure 29: Accuracy of Type V Companies in Consulting

I Analysis of Reasoning Failures

Model	Error Type	Description
Claude-3.7-Sonnet	Floating Points Error	Incorrect rounding in Accounts Receivable
	Reasoning Consistency	Inconsistent cash balance calculation
	Methodology Chain Breaking	Missed bank transfer in cash calculation
DeepSeek-V3	Factual Deviation	Added non-existent payable interest
	Knowledge Retrieval Deficiency	Failed to identify audit date error
	Floating Points Error	Lost precision in Taxes Payable
GPT-4.1	Methodology Chain Breaking	Ignored depreciation in cash flow
	Critical Data Omission	Omitted prepaid expenses in assets
GPT-4.1-mini	Contextual Inconsistency	Misclassified audit opinion type
	Logical Calculation Error	Inventory turnover ratio reversed
	Multi-Step Calculation Error	Skipped Account Receivabl in Total Assets End Value calculation
GPT-4o-mini	Factual Deviation	Misreported Account Payable as equity
	Format Handling	Failed parsing complex raw transaction data
GPT-4.1-nano	Logical Calculation Error	Fix asset purchase misclassified as cash inflow
	Arithmetic Error	Trial balance summation error
	Format Handling	Failed to perform cross financial statement analysis

Table 36: Comparative Analysis of Critical Errors Across Large Language Models

J Costs

The costs of LLMs for running all the tasks once are calculated based on the input token and completion token counts with their corresponding prices, which are shown in Table 37.

Model	Prompt	Completion	Cost
GPT-4o-mini	($0.15/MTok)	($0.6/MTok)	$18.35
GPT-4.1	($2/MTok)	8/MTok)	$265.47
GPT-4.1-mini	($0.4/MTok)	($1.6/MTok)	$49.87
GPT-4.1-nano	($0.1/MTok)	($0.4/MTok)	$11.22
o3-mini	($1.1/MTok)	($4.4/MTok)	$186.94
Claude-3.7-Sonnet	($3/MTok)	($15/MTok)	$396.07
DeepSeek-V3-2503	(2RMB/MTok)	(8RMB/MTok)	239.27RMB

Table 37: Cost for LLMs