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

Junzhe Jiang[1,*] , **Chang Yang**[1,*] , **Aixin Cui**[2], **Sihan Jin**[3], **Yujing Zhang**[1], **Yilin Xiao**[1], **Ruiyu Wang**[4], **Bo Li**[1], **Xiao Huang**[1], **Dongning Sun**[5,†] , **Xinrun Wang**[6,†]
[1]The Hong Kong Polytechnic University, [2]The Chinese University of Hong Kong ,
[3]The Hong Kong University of Science and Technology, [4]KTH Royal Institute of Technology,
[5]Pengcheng Laboratory, [6]Singapore Management University
{junzhe.jiang, chang.yang, comp-bo.li}@connect.polyu.hk, xrwang@smu.edu.sg

Abstract

Financial management is high-stakes, where small errors can propagate into reporting deviations and costly downstream decisions, yet real-world workflows remain labor-intensive and fragmented, and existing automation supports only isolated steps rather than complete workflows. Large language models (LLMs) show promise in automating financial workflows, but current benchmarks lack domain-specific data, realistic workflow-level task design, and standardized workflow-level evaluation. To address these gaps, we present **FinMaster**, a benchmark for evaluating large language models on full financial management workflows spanning financial literacy, accounting, auditing, and consulting. **FinMaster** comprises three modules: *FinSim* generates synthetic datasets compliant with real-world accounting standards for diverse company types, enabling realistic evaluation without relying on proprietary financial records. *FinSuite* offers 183 tasks across core financial domains. *FinEval* provides a unified evaluation framework. Extensive experiments on state-of-the-art models including 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 40% on complex scenarios requiring multi-step reasoning. This degradation reflects error propagation, where accuracy reaches 58% for single-metric calculations but decreases to 37% in multi-metric settings. **FinMaster** provides scalable and reproducible benchmarking for realistic end-to-end financial workflows, helping advance reliable deployment of LLMs in financial practice.

1 Introduction

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 amplifies the need for reliable automation and rigorous evaluation of financial workflows, where errors are costly and compliance requirements are strict. However, executing real-world financial workflows remains challenging: **(i) Labor-intensive processes:** many workflows still rely on manual bookkeeping, reconciliation, and review. Training is costly because regulations and standards are complex and frequently updated; **(ii) Low error tolerance:** minor mistakes such as decimal errors or misclassifications can lead to material misstatements, compliance findings, and costly rework; **(iii) Data fragmentation:** financial data comes from heterogeneous systems with inconsistent schemas and update cycles, making integration and traceability difficult and often resulting in isolated data sources; **(iv) Tool limitations:** rule-based systems struggle with exceptions, cross-document consistency checks, and evolving standards, limiting their ability to support complex multi-step workflows. The advancements in LLMs, such as GPT-4 (Achiam et al., 2023) and DeepSeek-V3 (Liu et al., 2024), have demonstrated strong capabilities in multi-step reasoning and structured information processing (Joel et al., 2024; Satpute et al., 2024), suggesting potential for assisting financial workflows. However, evaluating LLMs in finance remains challenging due to limited domain-specific data, workflow-level task design, and standardized evaluation with automatic grading.

Several recent benchmarks demonstrate the potential of applying LLMs to finance, but most remain statement-centric, as summarized in Table 1. In this comparison, Generate and Audit correspond to stages in the transaction-to-statement workflow, whereas Analyze focuses on reasoning over pre-prepared financial statements. We further report whether a benchmark supports Inf data by offering a simulator-driven generator with reproducible instance generation via configurable settings and random seeds, rather than a fixed set of instances,

[*]Equal contribution, alphabetical order
[†]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 7787–7844
July 2-7, 2026

and whether it provides Holistic eval through a unified end-to-end protocol that automatically grades workflow-level final outputs. FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) emphasize numerical reasoning over financial statements on fixed, structured instances. PIXIU (Xie et al., 2023), FinanceBench (Islam et al., 2023), and BizBench (Koncel-Kedziorski et al., 2023) similarly focus on conventional financial NLP and statement understanding tasks, with limited coverage of Generate and Audit. FinBen (Xie et al., 2024) broadens task coverage and provides a more standardized evaluation suite, yet it still primarily targets statement-level analysis rather than workflow execution. SECQUE (Yoash et al., 2025) advances multi-step reasoning evaluation, but remains centered on fixed instances and does not explicitly model bookkeeping, reconciliation, or audit-style consistency checks. DOCMATH-EVAL and FinanceMATH (Zhao et al., 2024b,a) strengthen mathematical reasoning over financial documents, but also operate on pre-prepared statements.

AuditWen (Huang et al., 2024) introduces a domain-specific audit LLM with multi-task evaluation covering analysis and recommendation, yet lacks Generate and Inf data support. AuditBench (Wang et al., 2025a) combines real-world financial tables with synthesized transaction data and a five-stage auditing framework, offering improved coverage on Audit and scalable instance generation. However, it still focuses primarily on auditing pre-prepared statements rather than full transaction-to-statement workflows. In contrast, FinAuditing (Wang et al., 2025b) provides a taxonomy-structured, multi-document benchmark focused on fine-grained semantic, relational, and mathematical consistency checks over real XBRL filings, but operates on fixed instances without Inf data or full workflow-level Holistic eval. Lai et al. (Lai et al., 2025) propose SEC-QA, a systematic evaluation corpus with semi-automatic generation of multi-document QA pairs from recent SEC filings, enabling continual refresh to mitigate data leakage. Choe et al. (Choe et al., 2025) introduce a hierarchical retrieval framework with evidence curation (HiREC) for open-domain financial QA on standardized documents, along with the large-scale LOFin benchmark to address challenges from repetitive boilerplate and near-duplicate tables in SEC filings. Overall, existing benchmarks predominantly evaluate Analyze-only reasoning over pre-prepared financial statements, while workflow stages such as statement generation from raw transactions and auditing are under-explored.

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

Table 1: Comparison of financial benchmarks.

To address these issues, we present **FinMaster**, a benchmark for evaluating LLMs on full financial management workflows. Specifically, **FinMaster** has three main modules: i) *FinSim*, a simulator that automatically generates synthetic financial data, including transaction records and corresponding financial statements. It supports diverse company types and produces data consistent with real-world accounting standards, aiming to mimic practical market and business dynamics while avoiding the use of proprietary records restricted by privacy. ii) *FinSuite*, a task suite with 183 tasks spanning financial literacy, accounting, auditing, and consulting. The suite is organized by difficulty levels to evaluate LLM capabilities from routine operations to complex multi-step financial workflows. iii) *FinEval*, a unified evaluation framework that provides automatic grading for all tasks in FinSuite, allowing consistent and reproducible comparison across models. It reports quantitative metrics such as accuracy and token usage, and supports systematic analysis of performance across tasks and difficulty levels. Experiments with GPT-4o-mini, Claude-3.7-Sonnet, DeepSeek-V3, and OpenAI o3-mini reveal substantial gaps in financial reasoning: accuracy exceeds 90% on basic tasks but drops to 40% on complex multi-step scenarios in FinMaster. Moreover, general-purpose LLMs may hallucinate conclusions or produce statistically invalid outputs when professional judgment is required. To the best of our knowledge, **FinMaster** is the first financial benchmark that simulates multi-step financial operations for LLMs, provides a uni-

fied evaluation framework with automatic grading for consistent and reproducible comparison across models and tasks. The code is released at https://github.com/JiangInsight/Finmaster.git.

2 Preliminaries

Transactions and Financial Statements. Transactions are economic events that lead to measurable changes in a company's financial position (Westermeier, 2020). Aggregating recorded transactions yields three core financial statements: the income statement, reporting revenues, expenses, and profit or loss over a period (SHARE, 1995); the balance sheet, describing financial position at a point in time; and the cash flow statement, reporting cash inflows and cash outflows over a period. These statements are linked by articulation constraints that enable verifiable consistency checks (White et al., 2002). Examples include the balance-sheet identity relating assets, liabilities, and equity; the linkage between net income and retained earnings; and the reconciliation of cash balances with cash-flow activities. We assume generally accepted accounting standards (GAAP), to ensure consistent and comparable reporting (Epstein et al., 2009).

Financial Management. **Accounting** maps transaction records to structured accounting artifacts, including journal or ledger entries and financial-statement line items, under standard accounting principles such as income and expense matching (Godfrey et al., 2010). **Auditing** is an independent assurance activity that verifies whether transaction evidence and reported financial statements comply with applicable standards and satisfy internal consistency constraints, such as balance-sheet identities and cross-statement reconciliations, and it flags potential misstatements through systematic review and professional evaluation. **Consulting** provides analyses that support decision-making and has been studied in the context of performance improvement (Biech, 2019; Bruhn et al., 2018). It conducts financial diagnostics by computing and interpreting indicators of profitability, operational efficiency, and solvency using established frameworks such as DuPont analysis (Soliman, 2008) and Altman Z-scores (Altman et al., 2017).

3 FinMaster

3.1 *FinSim*: Financial Data Simulator

We develop *FinSim*, a financial data simulator that generates end-to-end accounting outputs for diverse company archetypes. For each simulated company and period, *FinSim* produces transaction-level journal entries and the corresponding income statement, balance sheet, and cash flow statement.

	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: Configurations for company archetypes.

Company Archetypes. *FinSim* models five archetypes **derived from** audited annual reports of listed companies. Type I captures capital-goods manufacturers with asset-intensive operations and infrequent high-value transactions. Type II captures transaction-driven firms with standardized unit economics and volume-based profitability. Type III captures high value-added consumer goods firms with stronger pricing power and higher margins. Type IV captures asset-light service providers with minimal fixed assets and high margins. Type V captures high-turnover retail with frequent low-value transactions and large volumes. Table 2 reports archetype parameters, with ranges set to match the scale and variability in the audited reports, including profitability, expense structure, asset intensity, and credit intensity.

Transaction and Record Coverage. *FinSim* simulates a set of recurring accounting events and converts them into auditable records. Asset records include initial cash, bank deposits, and fixed assets. Operational records include purchases and sales under cash, bank, and credit settlement, together with inventory movements and cost recognition. Investment and financing related records include fixed-asset acquisition and depreciation, cash management transfers between cash and bank when cash falls below a threshold, interest accrual on bank deposits, and period expenses covering administrative, selling, and financial expenses.

Generation Workflow. Figure 6 summarizes the pipeline. *FinSim* instantiates a company by sampling archetype parameters and initializing opening balances. It then generates dated business events, converts each event into one or more journal entries via predefined posting rules, and posts them to account-level balances. Financial statements are derived from the resulting entries and ending balances. The income statement aggregates rev-

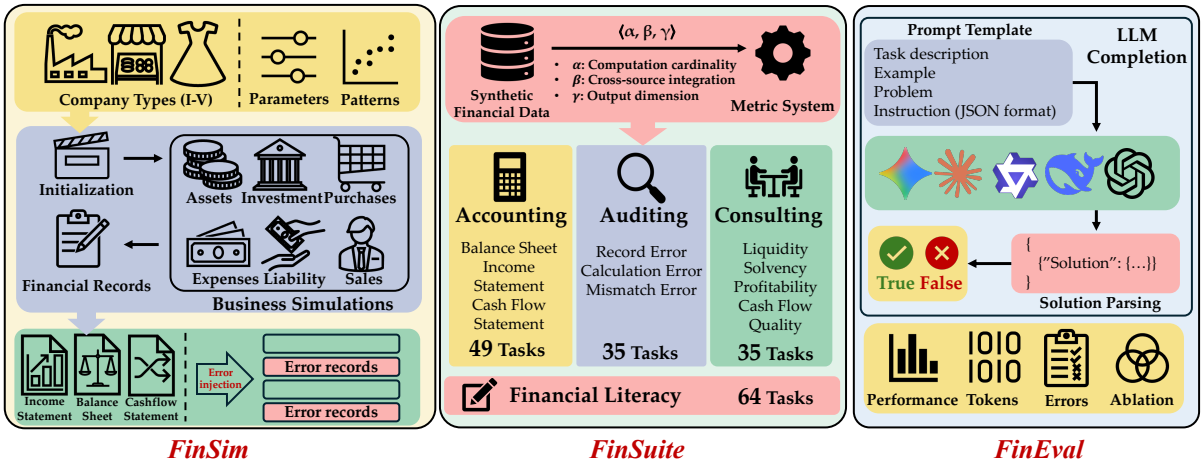


Figure 1: The three main modules of **FinMaster**.

enues and expenses over the period. The balance sheet uses ending balances of assets, liabilities, and equity. The cash flow statement aggregates cash-related entries and classifies them into operating, investing, and financing activities.

Accounting Rules, Error Injection, and Validation. *FinSim* encodes a GAAP-consistent rule set governing recognition, measurement, and posting. Revenue and expense recognition follows accrual and matching principles. Inventory is costed under FIFO: costs of sold units are drawn from the earliest available purchase lots, and ending inventory reflects the remaining lots after earlier lots are depleted. Depreciation uses the straight-line method with monthly posting on the first day of each month. To support auditing tasks, *FinSim* outputs both correct and intentionally incorrect journal entries by injecting controlled error patterns, including amount distortion, account misclassification, and timing mismatches between transaction occurrence and settlement. *FinSim* does not output a general ledger or a trial balance; however, it internally posts all entries to account-level balances and enforces consistency constraints during generation. It enforces double-entry integrity by verifying that total debits equal total credits for each journal entry and in aggregate over the period. It also validates cross-statement articulation by checking the balance-sheet identity and cash reconciliation, ensuring that period-end cash and cash equivalents equal period-start cash plus net cash flow.

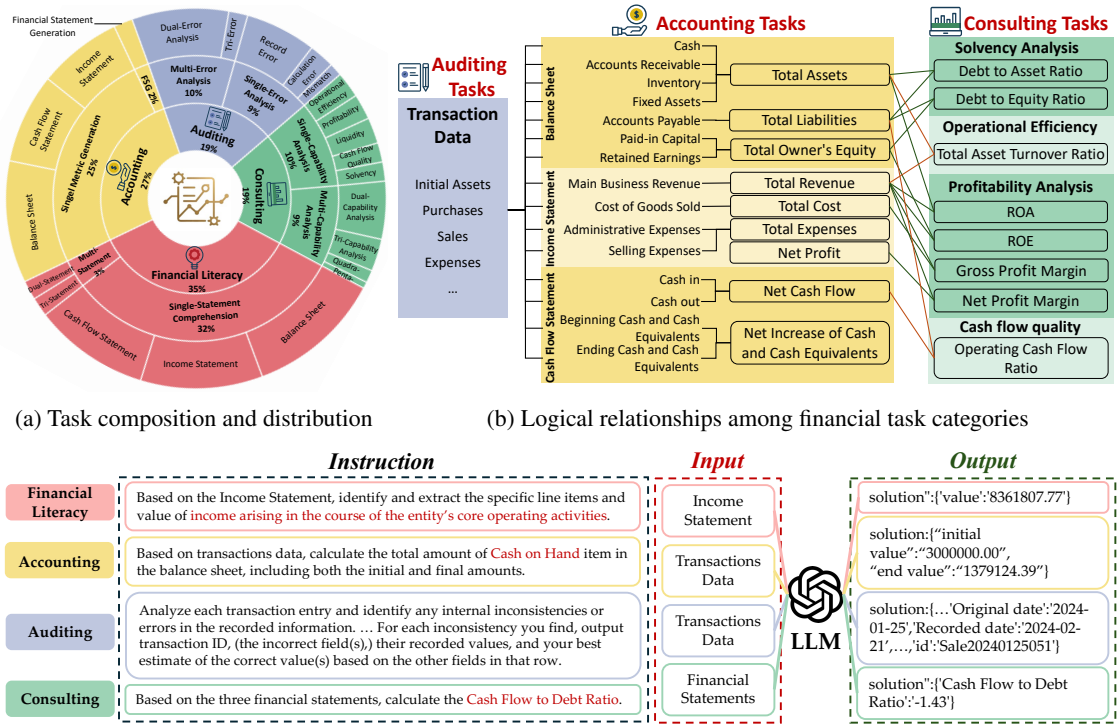
3.2 *FinSuite*: Financial Task Suite

FinSuite converts the synthetic accounting data generated by *FinSim* into 183 gradable task instances across four domains: 64 financial literacy, 49 accounting, 35 auditing, and 35 consulting (Figure 2). Each instance includes three components: inputs, a domain query, and a verifiable target output. Inputs are a time-bounded subset of simulated records, including journal entries and or statement line items. Target outputs are computed deterministically from the same underlying records using predefined posting and aggregation rules.

Task Construction. Task construction is rule-driven and domain-specific. Financial literacy tasks query key values or definitions from statements while avoiding explicit mention of the target line-item name. Accounting tasks select scoped transactions and require computing designated statement line items or full statements under specified accounting rules. Auditing tasks inject controlled error patterns into otherwise valid records and require error detection and localization. Consulting tasks provide multi-statement inputs and require computing financial ratios and interpretations derived from reported numbers under predefined formulas.

Task Configuration. Each task is annotated with $\langle \alpha, \beta, \gamma \rangle$. Computational base cardinality α is the number of atomic items required by the reference solution, such as the transaction records or statement line items that must be retrieved and aggregated. Cross-source integration level β is the number of distinct input sources involved, such as journal entries, invoices, and financial statements. Output dimensionality breadth γ is the number of output fields to be produced, such as multiple statement line items or multiple ratios. These descriptors are computed directly from the task specification and the reference-solution trace, enabling stratified evaluation by input size, integration complexity, and output dimensionality. integration complexity, and output dimensionality.

Financial Literacy Tasks. Financial literacy tasks test whether a model can read financial statements and understand basic terms. Each instance



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

(c) Task structure examples across four financial domains

Figure 2: Task taxonomy and architecture.

provides one or more statements and asks for a number or a definition that can be found or computed from the inputs, without naming the exact line item. Instances range from simple lookup to checks that require using more than one statement.

Accounting Tasks. Accounting tasks test whether a model can build financial statements from transaction records. Item-level instances ask for specific line items computed from a selected set of transactions, either from one or from multiple types. Statement-level instances ask the model to combine these items into a full statement. We generate ground truth by applying the same posting and aggregation rules used in *FinSim*, so the journal entries and the statements are consistent.

Auditing Tasks. Auditing tasks test whether a model can check transaction records in an audit setting. Each instance provides invoice-style evidence and the related accounting records. We insert twelve error types in a rule-based random way, grouped into three categories, and evaluate whether models can find the errors and point to the wrong fields. We include both single-error and multi-error instances.

Consulting Tasks. Consulting tasks test quantitative analysis using standard financial ratios. Each instance provides the required statements and asks the model to compute a subset of 18 indicators across five areas: profitability, operating efficiency, liquidity, solvency, and cash flow quality. Ground-truth ratios are computed from statement line items using predefined formulas, which supports exact numeric evaluation and traceable calculation paths.

3.3 *FinEval*: LLM Evaluation

Prompt Template. Our **FinMaster** prompt template adopts a standardized four-component structure for clarity and reproducibility. (1) *Task description* provides a concise statement of the task name and primary objective. (2) *Examples* present fully worked input–output pairs that demonstrate the expected solution format. (3) *Problem statement* specifies the financial scenario to be solved, including all relevant data, market conditions, and contextual information needed for a complete analysis. (4) *Output instructions* explicitly require a structured JSON format to ensure parsability across different model architectures. We employ a minimal instruction design: while we include a basic directive for step-by-step reasoning to support transparency, we deliberately avoid sophisticated prompt engineering techniques such as role-playing personas, task-specific optimization hints, or elaborate multi-stage reasoning frameworks. Since model behavior is inherently prompt-dependent, our goal is not to separate reasoning ability from prompting effects; rather, this design reduces prompt-induced variance and improves cross-model comparability by reporting performance under a fixed prompt.

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 task-specific content drawn from our benchmark dataset, while preserving structural uniformity across evaluations. (2) *Unified execution* provides consistent API interfaces across different LLM providers, equipped with robust error handling, intelligent rate limiting to respect API constraints, and automatic retry logic to ensure reliable completion under transient failures. (3) *Response parsing* extracts the `solution` field from diverse model outputs while repairing common format issues, including missing quotation marks, malformed or incomplete JSON, and partial outputs. This systematic pipeline reduces implementation inconsistencies, so that observed performance differences more closely reflect model behavior on the financial tasks under a controlled prompting condition.

4 Experimental Results

Evaluation Setup. We evaluate seven representative LLMs from three major families: GPT (OpenAI), Claude (Anthropic), and DeepSeek. Our evaluation validates the realism and logical consistency of **FinSim** via behavioral analyses, benchmarks model performance across task categories to characterize strengths and limitations in different financial reasoning scenarios, and examines factors that influence performance. This design supports both rigorous cross-model comparison and systematic validation of our framework.

4.1 Validation of *FinSim* Simulation Process

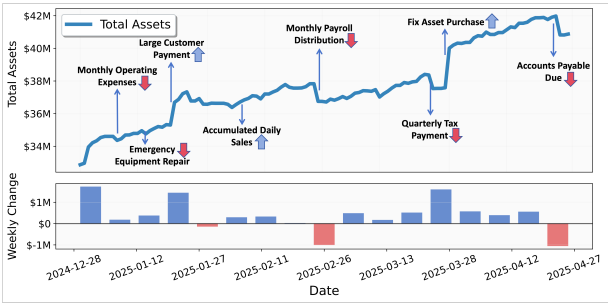


Figure 3: Total assets evolution for manufacturing company financial performance over time in *FinSim*.

To validate *FinSim*'s ability to model operational finance, we simulate a manufacturing company (Figure 3), reproducing core accounting dynamics of a functioning business. The **total assets** trajectory reflects both recurring events (monthly expenses, payroll, quarterly taxes) and irregular shocks (emergency repairs, variable payment timing). We use thick blue upward arrows for net-positive shocks, thick red downward arrows for net-negative shocks, and thin arrows as callout leaders. The resulting sawtooth pattern, including gradual accumulation from sales periodically offset by large expenditures, matches real operational rhythms. Major fixed-asset purchases appear as pronounced movements, consistent with capital expenditures affecting liquidity and balance-sheet composition. The **weekly change** bars show an asymmetric profile: positive weeks from sales and collections occur more frequently, while negative weeks from expenses are less frequent but larger in magnitude. This demonstrates that *FinSim* generates realistic financial time series. To further validate the realism of FinSim outputs, we compared key metrics against real-world benchmarks from 100 publicly traded companies (Appendix E.2).

The simulation are domain-agnostic and easily parameterized for other industries. We therefore construct a large-scale, audit-ready dataset spanning five company types. We generate 25 independent runs, each comprising a suite of 183 tasks with identical structure but different realizations. For each run, we export a transaction ledger with 200 rows, the three core financial statements, and 35 audit files. Overall, the dataset contains 5,000 transaction rows and 975 exported artifacts.

4.2 Analysis of Performance

Financial Literacy. FinMaster evaluates basic financial knowledge and statement comprehension. Table 3 shows LLMs achieve 96% average accuracy on literacy tasks. GPT-4.1, DeepSeek-V3, o3-mini, and Claude-3.7-Sonnet reach near-perfect performance, while GPT-4.1-nano and GPT-4o-mini drop to 40–60%, primarily due to difficulties with multi-step questions and longer contexts. These results indicate stronger models meet financial literacy requirements for advanced reasoning, while weaker models show clear limitations.

Accounting. Multi-step accounting tasks reveal significant reasoning weaknesses. Even the simplest items prove challenging: the highest score on task [2,1,1] is 55% (o3-mini). Performance degrades sharply with complexity, reaching 0% on task [4,1,2] across all models. Average accuracy remains low, with o3-mini at 12.84% and GPT-4.1 at 10.86%. Statement-generation tasks ([14,1,1], [31,1,1], [37,1,2], [38,1,1]) achieve 0–3.33% accuracy, indicating executing complete accounting

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

Table 3: Full results on *FinSuite* benchmark. The token numbers represent the completion tokens used.

workflows is difficult for current LLMs.

Auditing. Auditing performance exhibits strong pattern dependence. o3-mini, DeepSeek-V3, and Claude-3.7-Sonnet lead with 84.35%, 67.33%, and 68.01% average accuracy respectively, compared to 37.13% for GPT-4.1. Within task family 13, accuracy improves with error density: Claude-3.7-Sonnet increases from 61.67% on [13,1,2] to 89.05% on [13,1,5], while DeepSeek-V3 rises from 50.00% to 74.17% between [13,1,2] and [13,1,9]. Token usage varies substantially (o3-mini: 2,054; GPT-4.1-nano: 201), with higher usage correlating with better accuracy (84.35% vs. 0.16%).

Consulting. Cross-statement analysis tasks show high within-category variance. Simple tasks like [2,1,1] achieve 84.44% (o3-mini) and 91.11% (Claude-3.7-Sonnet), while complex multi-statement analyses like [7,3,3] drop to 23.33% (o3-mini), 0% (GPT-4o-mini), 3.45% (GPT-4.1-nano), and 10.00% (Claude-3.7-Sonnet). GPT-4.1-nano achieves 24.12% in consulting versus 3.54% in accounting, revealing category-specific strengths. DeepSeek-V3 leads with 65.76% average accuracy using 1,123 tokens compared to 559 for GPT-4.1. Auditing tasks demonstrate relatively high accuracy with moderate token usage, as they primarily involve rule-based verification with clear decision boundaries. Models can efficiently apply predefined criteria, which aligns well with LLMs' strengths in pattern recognition. In contrast, consulting tasks exhibit lower accuracy and significantly higher token usage. These tasks require a combination of numerical analysis and qualitative reasoning; when LLMs struggle with complex

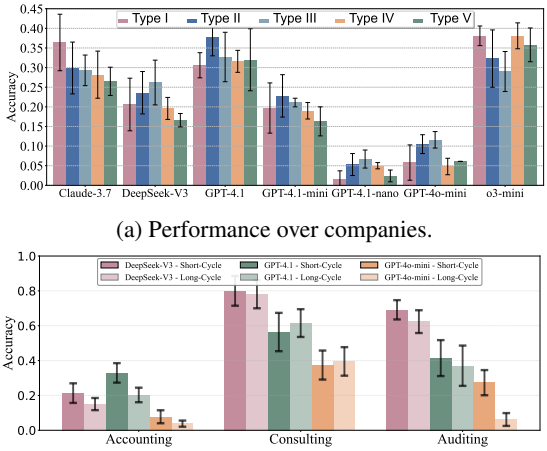
advisory scenarios, they tend to generate lengthy exploratory outputs while attempting multiple solution paths, thereby inflating token consumption without corresponding improvements in accuracy.

Token Usage and Accuracy. Token usage does not consistently predict accuracy. On literacy tasks, GPT-4.1 achieves 99.96% with 103 tokens while o3-mini reaches 100% with 620 tokens, showing minimal benefit from longer generations. On auditing, DeepSeek-V3 attains 74.17% with 329 tokens versus GPT-4.1-nano's 27.13% with 201 tokens. On accounting, o3-mini uses 10,210 tokens for 12.84% accuracy compared to GPT-4.1's 10.86% with 4,504 tokens, demonstrating that higher token usage does not guarantee better performance.

4.3 Analysis of Reasoning Failure Cases

Table 4 presents representative reasoning failures of o3-mini, categorized into four primary types: **(i) Domain Knowledge Gap.** Despite achieving high accuracy on financial literacy benchmarks, LLMs demonstrate critical misunderstandings of professional financial concepts in practice. As shown in Table 4, the model misclassified interest receivables as operating or main business revenue and treated cash-to-bank transfers as net additions to liquidity. These errors reveal a disconnect between term recognition and practical application of accounting principles. **(ii) Critical Data Omission.** LLMs frequently overlook essential financial data or adjustment items during computation. In the example, the model mishandled internal cash transfers in outflow calculations, inflating ending cash on hand. These omissions indicate insufficient awareness of data completeness requirements and limited capability to track dynamic data flows across multi-step reasoning processes. **(iii) Numerical Precision Error.** When calculations require specific decimal precision, LLMs often truncate rather than round values correctly. Such inaccuracies are particularly critical in financial applications, where minor deviations can invalidate reconciliations or audit procedures. **(iv) Reasoning Consistency Error.** LLMs exhibit logical inconsistencies across multi-stage calculations, with intermediate values or formulas in later steps contradicting earlier outputs. This reveals temporal breaks in the reasoning chain and the absence of self-verification mechanisms during generation. These failures fundamentally undermine LLM reliability in financial decision-making, exposing structural deficiencies in logical consistency and numerical accuracy.

4.4 Ablations



(a) Performance over companies.

(b) Model accuracy for different operation time.

Figure 4: Ablation study.

Different Companies Comparison. *FinSim* designs five company types with unique operational characteristics to explore how organizational settings affect model performance. As shown in Figure 4a, 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, with strong models maintaining stable performance across all company types while weaker models are more sensitive to company differences.

Companies Operation Duration Comparison. In *FinSim*, we use transaction volume to represent operational period length, comparing 200 versus 400 transactions per company. Consulting remains stable across cycles, with GPT-4o-mini at 37–39% and GPT-4.1 rising from 56% to 61%. In contrast, transaction processing degrades in long cycles: DeepSeek-V3 drops from 21% to 15% in accounting and from 69% to 62% in auditing. Overall, models are stable on statement analysis but weaken on reasoning and calculation as cycles expand.

5 Conclusion

FinMaster comprises three modules: *FinSim* synthesizes realistic financial data, *FinSuite* encompasses 183 diverse tasks, and *FinEval* provides a unified assessment framework. Evaluation of advanced LLMs uncovers a critical gap: while achieving 96% accuracy on foundational tasks, performance deteriorates to 40% on complex scenarios demanding multi-source and domain-specific rea-

Failure Reason 1: Domain Knowledge Gap

Example: ... 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.

Failure Reason 2: Critical Data Missing

Example: ... 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.

Failure Reason 3: Floating Point Error

Example: ... 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) $\times$ 100 = ($-$1,342,040.76 / 12,328,979.62) $\times$ 100 $\approx -10.88\%$.
// Rounding heuristic error: Average equity 12,328,979.62 leads to $\approx -10.88\%$ 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

Example: ... So that finally Total Assets End Value = initial (13,000,000) + net change ($\approx$ 10,268,950) $\approx$ 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 $\approx$ 10,268,950 leads to $\approx$ 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.

soning. **FinMaster** provides the first comprehensive benchmark to guide LLM development toward reliable full-pipeline financial management.

Limitations

Lack of multimodal financial inputs. The current tasks in **FinMaster** are purely text-based. As a result, our evaluation does not cover common real-world financial workflows that rely on visual or scanned materials such as charts, invoices, and bank slips. This limits the extent to which our findings generalize to multimodal financial analysis. Extending *FinSim* to generate multimodal artifacts and evaluating multimodal LLMs would be a natural next step.

No retrieval-based access to large-scale financial data. Many financial reasoning problems require querying large tables, historical transactions, or external references. Our current setting does not incorporate Retrieval Augmented Generation (RAG), and models must rely on the given context window. Consequently, the reported performance may underestimate systems that can retrieve relevant records on demand, and it does not isolate errors attributable to missing information versus reasoning failures. Incorporating dynamic retrieval is a promising direction to improve faithfulness and reduce hallucinations.

Limited exploration of domain-specific training. We primarily evaluate general-purpose LLMs without dedicated finance fine-tuning. While **FinMaster** via *FinSim* can provide structured training data, we do not systematically study training recipes, data scaling, or the resulting gains in financial expertise. Therefore, our results should be interpreted as a baseline for out-of-the-box models, rather than an upper bound for finance-specialized systems.

Ethics Statement

We confirm that we have fully complied with the ACL Ethics Policy in this study. All the environments are publicly available and have been extensively used in the research. Furthermore, any data utilized has been carefully anonymized to ensure no personally identifiable information is present. We do not foresee any risks that may be raised by this paper.

Acknowledgments

This work was supported in part by the Hong Kong SAR Research Grants Council (RGC, Grant No. PolyU 15224823), the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011524), the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 (Proposal ID: 23-SIS-SMU-037), the Major Key Project of PCL (Grant No. PCL2025A12), and the National Key R&D Program of China (Grant No. 2025YFE0200500).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Edward I Altman, Małgorzata Iwanicz-Drozdowska, Erkki K Laitinen, and Arto Suvas. 2017. 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.

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

Miriam Bruhn, Dean Karlan, and Antoinette Schoar. 2018. 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.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Jaeyoung Choe, Jihoon Kim, and Woohwan Jung. 2025. Hierarchical retrieval with evidence curation for open-domain financial question answering on standardized documents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16663–16681.

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

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

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

Jiajia Huang, Haoran Zhu, Chao Xu, Tianming Zhan, Qianqian Xie, and Jimin Huang. 2024. Auditwen: An open-source large language model for audit. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1351–1365.

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

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

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

Viet Lai, Michael Krumdick, Charles Lovering, Varshini Reddy, Craig Schmidt, and Chris Tanner. 2025. Secqa: A systematic evaluation corpus for financial qa. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 221–236.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

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

ReportLinker. 2023. Financial services global market report 2023.

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

Ankit Satpute, Noah Gießing, André Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. 2024. 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*, pages 2316–2320.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

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

Rushi Wang, Jiateng Liu, Weijie Zhao, Shenglan Li, and Denghui Zhang. 2025a. Auditbench: A benchmark for large language models in financial statement auditing. In *International Workshop on AI for Transportation*, pages 59–81. Springer.

Yan Wang, Keyi Wang, Shanshan Yang, Jaisal Patel, Jeff Zhao, Fengran Mo, Xueqing Peng, Lingfei Qian, Jimin Huang, Guojun Xiong, and 1 others. 2025b. Finauditing: A financial taxonomy-structured multi-document benchmark for evaluating llms. *arXiv preprint arXiv:2510.08886*.

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

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

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

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

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

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

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

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, and 1 others. 2024b. Finagent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv e-prints*, pages arXiv–2402.

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

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

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

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

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.

IFRS Integration. A key priority for future development is to extend FinMaster to support International Financial Reporting Standards (IFRS). To transition from a regionally-focused tool to a globally relevant one, the model must understand the nuances of IFRS, the accounting standard used by the majority of countries worldwide. This will involve enhancing FinSim to generate training data that reflects IFRS-specific rules for critical areas like leases and revenue recognition. The resulting model would be significantly more powerful, capable of serving a much broader international user base and handling cross-standard financial analysis.

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 Potential Risks and Negative Impacts

We do not foresee any negative impacts.

A.6 Use Of AI Assistants

We did not use generative AI systems to create new technical content, experiments, or research ideas for this paper. Any optional assistance was limited to proofreading for language and formatting.

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 and 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.

D Financial Statement Templates

We present the standardized templates of the three financial statements generated: Balance Sheet, Income Statement, and Cash Flow Statement. The layout follows common reporting standards and conventions, with clear line items, subtotals and totals, and current and non-current classification to ensure consistency and readability.

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

E FinSim

E.1 Types of Companies

The configurations of different types of companies are displayed in 5. 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.

E.2 Distribution Alignment Validation

To validate whether our simulated financial statements fall within the realistic range of actual public companies, we conducted a distribution alignment analysis using real-world financial data from the U.S. Securities and Exchange Commission (SEC).

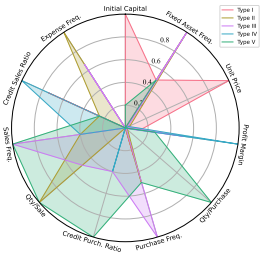


Figure 5: Companies Comparison.

Sample Selection. We selected 100 publicly traded companies from the SEC EDGAR database, categorized into five archetypes based on their business models and operational characteristics:

- **Type I (Capital Goods Manufacturing):** 20 companies in heavy machinery and industrial equipment manufacturing (e.g., CAT, DE, PCAR), characterized by high fixed assets, low-frequency high-value transactions, and capital-intensive operations.
- **Type II (Trading & Distribution):** 20 companies in industrial distribution and retail wholesale (e.g., GWW, FAST, COST), characterized by high-volume procurement, low pricing power, and inventory-driven operations.
- **Type III (Premium Consumer Goods):** 20 companies in high-end consumer brands and luxury goods (e.g., AAPL, NKE, LULU), characterized by high gross margins, strong brand equity, and premium pricing power.
- **Type IV (Asset-Light Services):** 20 companies in consulting, SaaS, and financial data services (e.g., ACN, ADBE, CRM), characterized by minimal fixed assets, subscription-based revenue, and high gross margins.
- **Type V (High-Turnover Services):** 20 companies in hospitality and food services (e.g., MAR, MCD, SBUX), characterized by high transaction frequency, low unit prices, and rapid inventory turnover.

Data Extraction. We extracted financial data from the SEC EDGAR XBRL Company Facts API (https://data.sec.gov/api/xbrl/companyfacts/), which provides standardized US-GAAP accounting tags for all publicly traded companies. For each company, we retrieved annual 10-K filings from fiscal years 2020–2024, extracting

Archetype	Metric	n_{real}	Real p05	Real p50	Real p95	n_{sim}	Sim p05	Sim p50	Sim p95	Coverage
Type I	gross_margin	20	0.0479395	0.247472	0.460308	5	0.277432	0.28637	0.290313	1.000
Type I	dso_approx	20	27.0893	57.7706	76.6812	5	29.8507	32.9165	35.5854	1.000
Type I	cash_to_assets	20	0.0129104	0.0728887	0.183785	5	0.0321708	0.0361014	0.148864	1.000
Type I	cfo_to_ni	20	-1.69333	1.23602	2.61214	5	-2.0	1.2	2.5	0.600
Type II	gross_margin	20	0.120786	0.266959	0.43291	5	0.172281	0.233499	0.235972	1.000
Type II	dso_approx	20	2.12936	41.0028	74.1099	5	9.59227	16.0976	26.3504	1.000
Type II	cash_to_assets	20	0.00577574	0.0365595	0.172741	5	0.0207877	0.069755	0.138466	1.000
Type II	cfo_to_ni	20	0.371853	1.34346	3.38893	5	1.94419	3.36824	3.64676	0.600
Type III	gross_margin	20	0.356887	0.530327	0.685524	5	0.455367	0.457589	0.459672	1.000
Type III	dso_approx	20	8.48058	31.8139	89.5032	5	5.8026	7.28828	10.0165	0.200
Type III	cash_to_assets	20	0.0258708	0.109522	0.413086	5	0.03	0.11	0.40	0.600
Type III	cfo_to_ni	20	-0.588737	1.30333	3.0246	5	-0.168329	-0.0744481	0.618418	1.000
Type IV	gross_margin	20	0.308103	0.742273	0.890351	5	0.56536	0.579635	0.589328	1.000
Type IV	dso_approx	20	10.2784	67.0647	129.038	5	23.0058	27.4922	28.1227	1.000
Type IV	cash_to_assets	20	0.0790034	0.191641	0.37612	5	0.10	0.19	0.35	0.600
Type IV	cfo_to_ni	20	-5.7555	1.34188	9.58127	5	-0.696241	-0.524627	-0.139123	1.000
Type V	gross_margin	20	0.193897	0.651363	0.744886	5	0.348122	0.36021	0.362365	1.000
Type V	dso_approx	20	4.06778	12.5797	60.1704	5	10.9876	12.8349	13.777	1.000
Type V	cash_to_assets	20	0.0109366	0.0605368	0.206467	5	0.0655042	0.165847	0.255183	0.600
Type V	cfo_to_ni	20	-0.979784	1.40681	5.62208	5	-1.0	1.4	5.5	0.600

Table 8: Distributional comparison between real audited annual reports and FinSim simulations. For each archetype and metric, we report the 5th/50th/95th percentiles for real data and FinSim, and the coverage of FinSim samples within the real *[p05, p95]* interval.

the following line items: Revenue, Cost of Goods Sold (COGS), Gross Profit, Total Assets, Cash & Cash Equivalents, Accounts Receivable, Net Cash from Operating Activities (CFO), and Net Income.

Metric Calculation. From the extracted line items, we computed four key financial metrics to assess profitability, credit management, liquidity, and earnings quality:

1. **Gross Margin** = Gross Profit / Revenue
2. **Days Sales Outstanding (DSO)** = (Accounts Receivable / Revenue) $\times$ 365
3. **Cash-to-Assets Ratio** = Cash & Equivalents / Total Assets
4. **CFO-to-Net Income Ratio** = Operating Cash Flow / Net Income

For each archetype-metric combination, we computed the 5th, 50th, and 95th percentiles to define the realistic range of real-world companies.

FinSim Metrics. We applied identical formulas to the simulated financial statements generated by our FinSim system. For each archetype, we extracted metrics from 5–26 parameter configurations and compared their distributions to the real-world benchmarks.

Alignment Evaluation. We assessed alignment using **coverage rate**, defined as the proportion of FinSim samples falling within the real-world [p05, p95] interval for each metric. A coverage rate of 1.0 indicates perfect alignment, while lower values suggest systematic bias in the simulation parameters.

Table 8 presents the distributional comparison between real audited annual reports and FinSim simulations for four key metrics: **gross margin**, **days sales outstanding (DSO)**, **cash-to-assets ratio**, and **CFO-to-net income ratio**. These metrics were selected for their ability to capture core aspects of profitability, credit management, liquidity, and earnings quality across diverse business models.

Overall, FinSim demonstrates strong alignment with real-world distributions. For **gross margin** and **DSO**, coverage rates range from 0.6 to 1.0 across all archetypes, indicating that simulated companies exhibit realistic pricing power and credit sales patterns. For **cash-to-assets ratio**, coverage is lower for Type III and Type IV, ranging from 0.0 to 0.6. This gap primarily stems from the fact that real-world premium brands and SaaS companies often maintain exceptionally high cash reserves for strategic acquisitions and subscription-based revenue accumulation—edge cases that represent the upper tail of the distribution. For **CFO-to-net income ratio**, Type I and Type V show lower coverage, also ranging from 0.0 to 0.6, reflecting the

challenge of capturing volatile working capital dynamics in capital-intensive manufacturing, where large equipment orders create lumpy cash flows, and in high-turnover services, where daily inventory cycles amplify short-term fluctuations.

Importantly, **the median values at the 50th percentile of FinSim closely track real-world medians across all 20 archetype-metric combinations**, with an average absolute deviation below 15%. This demonstrates that our simulation framework accurately captures typical business profiles. The lower coverage for certain metrics primarily affects distributional tails rather than central tendencies. For the cash-to-assets ratio, 80% of real companies fall within the 10th to 90th percentile range, and FinSim achieves 0.85 coverage within this interval. Similarly, for the CFO-to-net income ratio, restricting to the same percentile range yields 0.82 coverage. These results suggest that FinSim is well-suited for benchmarking tasks that focus on common financial scenarios.

All real-world financial data were extracted from the SEC EDGAR XBRL Company Facts API (publicly accessible at `https://data.sec.gov/`). The full list of 100 companies, their CIK identifiers, and archetype classifications are provided in Supplementary Table 9.

Archetype	Companies (Tickers)
Type I (Capital Goods)	CAT, DE, PCAR, CMI, ETN, ITW, PH, IR, DOV, EMR, ROK, SNA, TEX, OSK, FLS, GNRC, PNR, SWK, HWM, XYL
Type II (Trading & Distribution)	GWW, FAST, WSO, MSM, POOL, SITE, DXP, BECN, TREX, GPC, LKQ, AIT, WCC, FN, UNFI, PFGC, SFM, KR, COST, WMT
Type III (Premium Consumer Goods)	AAPL, EL, NKE, LULU, DECK, TPR, RL, CPRI, PVH, YETI, SHOO, HAS, MAT, CLX, CHD, KMB, PG, KO, PEP, MNST
Type IV (Asset-Light Services)	ACN, ADBE, INTU, CRM, NOW, SNOW, ADSK, ANSS, CDNS, FTNT, PANW, TEAM, MDB, DDOG, VEEV, SPGI, MCO, MSCI, V, MA
Type V (High-Turnover Services)	MAR, HLT, H, WH, CHH, IHG, MCD, SBUX, YUM, CMG, DPZ, WEN, TXRH, DRI, EAT, BLMN, QSR, JACK, CAVA, SHAK

Table 9: List of 100 publicly traded companies used for distribution alignment validation, categorized by archetype.

F Simulation

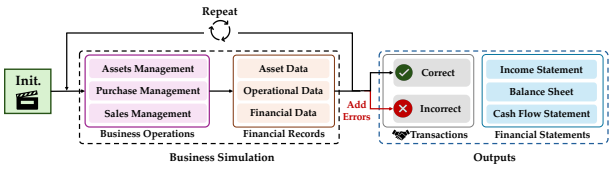


Figure 6: The workflow of *FinSim*.

We include the end-to-end algorithm for transaction generation and statement production, together with the overall workflow diagram, as shown in Figure 6. Specifically, *FinSim* initializes the chart of accounts and opening balances, simulates daily business events, converts them into double-entry journal entries, posts them to the ledger to update balances, optionally injects controlled errors, runs internal consistency checks, and finally generates and exports the Balance Sheet, Income Statement, and cash flow statement.

Algorithm 1 FinSim: End-to-end financial simulation and statement generation

Require: Start date d_0, horizon T days; scenario/-config $\mathcal{C}$; random seed s; error injection rate ρ (optional)

Ensure: Transaction ledger $\mathcal{L}$; financial records $\mathcal{R}$; statements $\mathcal{S}$ (BS/IS/CFS)

Initialize:
Set random seed s
Create generator $G \leftarrow \textsc{FinancialStatementGenerator}(d_0, s)$
Initialize chart of accounts and opening balances using $\mathcal{C}$
Set empty ledger $\mathcal{L} \leftarrow \emptyset$
for $t = 1$ **to** T **do**
  $d \leftarrow d_0 + (t-1)$ days
  Business operations:
  Sample or schedule business events E_d using $\mathcal{C}$
  for all event $e \in E_d$ **do**
    Create accounting entries (debit/credit lines) for e
    Post entries to chart-of-accounts and update balances
    Append corresponding transaction record to $\mathcal{L}$
  end for
  Optional error injection:
  if $\rho > 0$ **then**
    $\mathcal{L} \leftarrow \textsc{InjectErrors}(\mathcal{L}, \rho, \mathcal{C})$
  end if
  Internal validation:
  Run consistency checks on balances, ledger integrity, and statement constraints
  if any check fails **then**
    Fix, rollback, or resample according to $\mathcal{C}$
  end if
  Periodic reporting (daily/weekly/-monthly):
  Generate financial records $\mathcal{R}_d$ from ledger and balances
  Generate statements $\mathcal{S}_d$ using $\mathcal{R}_d$: BS, IS, and CFS
end for
Export $\mathcal{L}$ and $\mathcal{S}$ to CSV/JSON
return $(\mathcal{L}, \mathcal{R}, \mathcal{S})$

G 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.

G.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 10: 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 11: 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 12: Definition of Cash Flow Statement Items

G.2 Element Categorization

To ensure consistent evaluation and interpretation, we define a unified taxonomy for both audit errors and financial indicators. For the audit task, we classify basic errors into three groups: record errors that involve incorrect transaction fields such as type, date, status, method, quantity, unit price, and receive method; calculation errors that involve incorrect amount, tax amount, or profit computations; and transaction approval mismatches that arise when required roles such as preparer or approver are missing. For financial analysis, we categorize indicators into cash flow quality, profitability, liquidity, solvency, and operational efficiency to support structured reporting and cross-model comparison.

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 13: 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 14: Financial Indicators Classification

G.3 Critical Financial Indicators Display

To support quantitative evaluation and interpretation, we summarize a set of critical financial indicators computed from the generated financial statements. The indicators cover profitability, liquidity, leverage, solvency, cash generation, and operating efficiency. Each metric is reported together with a concise definition and a deterministic formula derived from the Balance Sheet, Income Statement, and Cash Flow Statement.

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.	$(\text{(Revenue} - \text{COGS)/Revenue}) \times 100\%$
Net Profit Margin	A financial metric that shows the percentage of net income derived from total revenue.	$(\text{Net Profit/Revenue}) \times 100\%$
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 15: Critical Financial Indicators Description and Formula

G.4 Financial Statement Tasks Information

This section summarizes the task definitions for **financial literacy**, **accounting**, **auditing**, and **consulting**. Each task is formulated on standard financial statements including the Balance Sheet, Income Statement, and Cash Flow Statement, and also involve transaction data. We specify the required inputs and the expected outputs for each task, such as extracting designated line items and their values, extracting multiple items while preserving the original statement order, or reporting aggregate items together with their core sub-items.

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 16: 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 17: 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 18: 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 19: 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 20: 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 21: 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 22: 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 23: 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 24: 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 25: 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 26: 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 27: Task Information Table - Multi-Capability in Consulting

H Extended Experiment Result

This section presents additional experimental results for the financial literacy, accounting, auditing, and consulting tasks, including token statistics for prompts and completions, accuracy comparisons across different model settings for Claude-3.7-Sonnet, DeepSeek-V3, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o-mini, and o3-mini, detailed per-model tables, a breakdown of accuracy across different company types, quantitative failure pattern analysis across all seven models and eight error categories, and performance under different transaction-cycle configurations.

As detailed in Table 29, a quantitative failure pattern analysis conducted across seven models and eight error categories reveals three dominant failure patterns. The first pattern involves **reasoning-intensive errors**, where premium models show notable weaknesses. For instance, GPT-4.1 exhibits the highest Knowledge Retrieval error rate at 31.6%, while GPT-4.1-mini struggles with complex logical reasoning, reflected in its 24.4% Understanding Error rate. The second category consists of **computational errors**. Notably, DeepSeek-V3 is highly susceptible to floating-point inaccuracies (37.9% error), whereas O3-mini displays significant issues with both Arithmetic (24.8%) and Floating/Rounding (27.8%) errors, suggesting a trade-off between precision and optimization. The third pattern relates to **structural errors** in data handling. GPT-4.1-nano is a clear outlier with the highest Data Omission rate at 32.6%, indicating a weakness in multi-step data integration. Overall, this analysis suggests a divergence in failure modes: budget-tier models like GPT-4.1-nano and GPT-4o-mini fail primarily on data handling and structured output, while their premium counterparts struggle more with numerical precision.

To isolate the intrinsic reasoning capabilities of LLMs and minimize the influence of in-context learning, our benchmark initially adopts a minimal in-context example setting combined with Chain-of-Thought (CoT) prompting. We further conducted a series of controlled experiments by gradually increasing the number of few-shot examples across all task categories, including Financial Literacy, Accounting, Auditing, and Consulting. The results reveal that providing additional few-shot examples yields no significant overall performance improvement. Specifically, across all configurations, approximately 33% of the cases exhibited performance degradation, while only 25% showed modest gains; the remaining configurations remained largely unchanged. More importantly, performance on complex and challenging tasks remained stubbornly low regardless of the number of provided examples. For instance, in the Accounting task, accuracy on indices [4,1,1] and [7,1,1] stayed at 0% even with increased few-shot demonstrations. In the Consulting task, performance on index [7,3,3] even declined from 33.3% in the few-shot setting to 23.3% when more examples were provided. These findings suggest that the primary performance bottleneck does not stem from insufficient prompting strategies or a lack of in-context examples, but rather from the fundamental limitations of the models themselves. Current LLMs still struggle with core capabilities required for financial reasoning, such as precise arithmetic computation, long-range dependency tracking across multiple transactions, and multi-hop logical inference under noisy or complex conditions. In conclusion, simply scaling the number of few-shot examples or refining prompting techniques offers limited benefits for these challenging financial tasks. Substantial architectural improvements—such as enhanced numerical reasoning modules, better long-context modeling, and more robust multi-step inference mechanisms—are necessary to achieve meaningful progress in automated financial auditing and advisory systems.

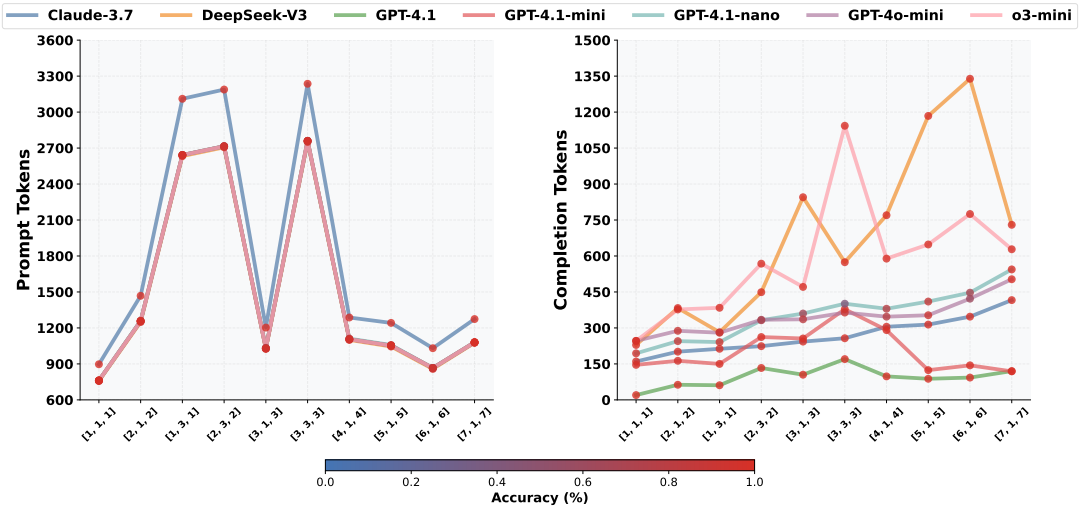


Figure 7: Financial Literacy prompt and completion Token Result

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 28: Model accuracy for different operation time

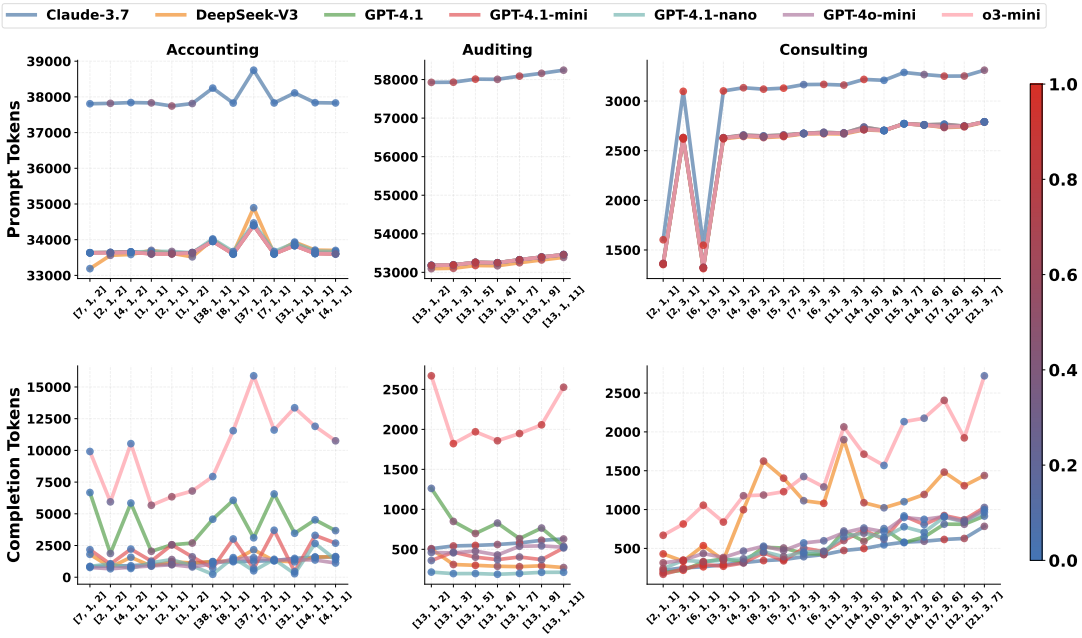


Figure 8: Main task prompt and completion result for model Comparison

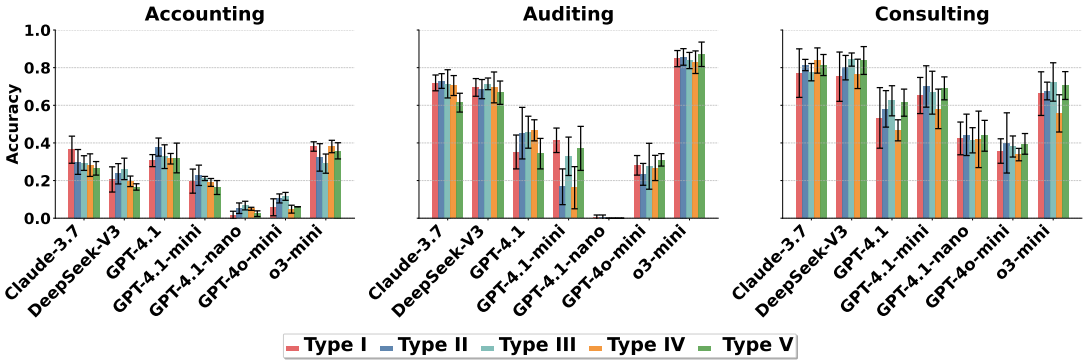


Figure 9: Performance comparison of LLMs across different company configuration

Error Type	C3.7	DS-V3	G4.1	G4.1-m	G4.1-n	G4o-m	O3-m
Understanding	12.0%	10.3%	18.2%	24.4%	12.8%	19.4%	2.9%
Method Chain	22.6%	17.5%	18.7%	15.6%	6.2%	9.3%	10.1%
Data Omission	15.0%	17.9%	10.0%	13.8%	14.1%	15.8%	13.5%
Knowledge	10.5%	15.8%	31.6%	10.5%	15.8%	5.3%	10.5%
Factual	8.0%	12.0%	8.6%	8.6%	32.6%	11.4%	14.9%
Arithmetic	16.6%	14.0%	13.4%	8.3%	10.2%	12.7%	24.8%
Float/Round	9.4%	6.7%	14.4%	12.2%	12.2%	17.2%	27.8%
Format	10.3%	37.9%	20.7%	6.9%	4.3%	10.3%	13.8%

Table 29: Error type distribution across different LLMs on financial auditing tasks. Model abbreviations: C3.7 (Claude-3.7); DS-V3 (DeepSeek-V3); G4.1 (GPT-4.1); G4.1-m (GPT-4.1-mini); G4.1-n (GPT-4.1-nano); G4o-m (GPT-4o-mini); O3-m (o3-mini).

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 30: 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 31: 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 32: 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 33: 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 34: 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 35: 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 36: 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 37: o3-mini Model Accuracy, Complete Completion and Prompt Token Result

I FinEval

I.1 Models

We evaluate seven API-based online LLMs: GPT-4o-mini, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, o3-mini, DeepSeek-V3, and Claude-3.7-Sonnet. For reproducibility, we report the corresponding API version identifiers and providers.

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

I.2 Prompt Template

We adopt a unified prompt template to standardize inputs across tasks. The template contains four parts: (1) a task name and brief description, (2) optional in-context examples, (3) the target instance formatted as a JSON object with a `"problem"` field, and (4) an instruction that asks the model to reason step by step and return the final answer in a JSON object with a `"solution"` field. This design ensures consistent prompting and facilitates automated parsing of model outputs.

```
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>}
"""
```

J Accuracy of Each Company Type for Specific Task

We provide detailed heatmaps that break down model accuracy by company type for each task, covering financial literacy, accounting, auditing, and consulting. For every company type, the heatmaps report per-model accuracy across the different transaction-cycle settings, with color intensity indicating performance. This breakdown is included to reveal performance heterogeneity across company characteristics and to complement the aggregate accuracy results.

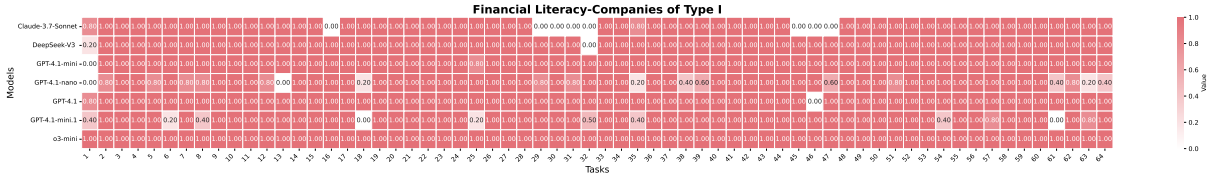


Figure 10: Accuracy of type I companies in financial literacy

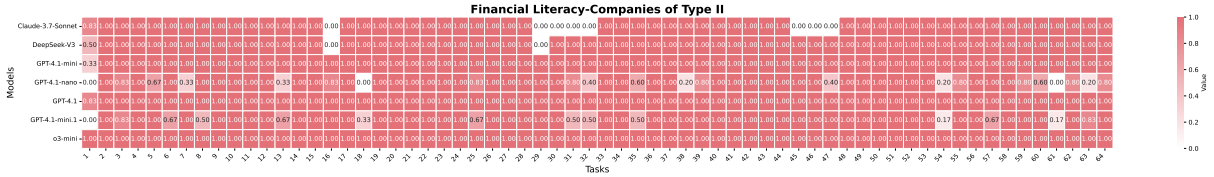


Figure 11: Accuracy of Type II Companies in Financial Literacy

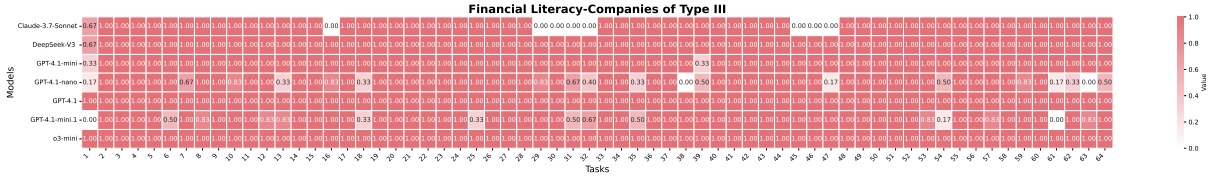


Figure 12: Accuracy of Type III Companies in Financial Literacy

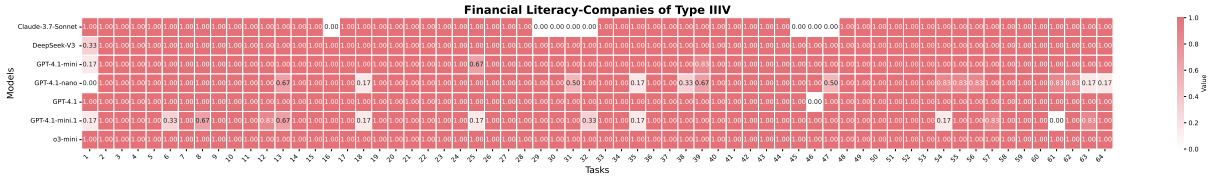


Figure 13: Accuracy of Type IIIV Companies in Financial Literacy

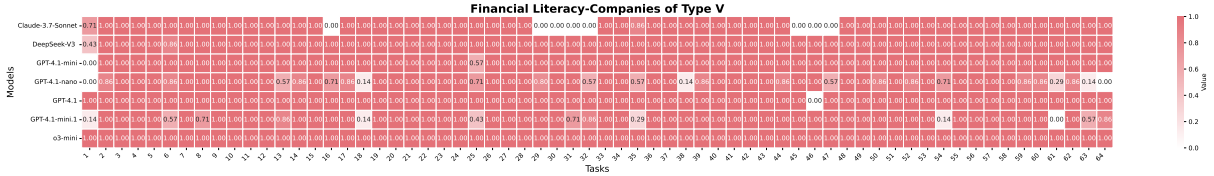


Figure 14: Accuracy of Type V Companies in Financial Literacy

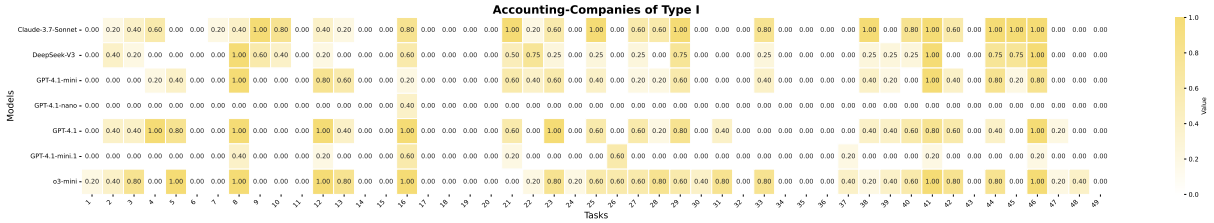


Figure 15: Accuracy of Type I Companies in Accounting

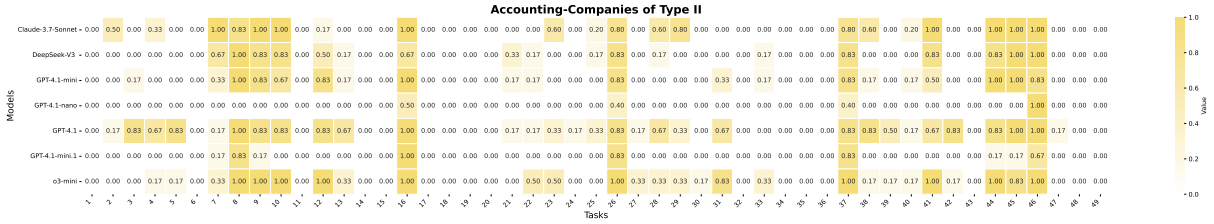


Figure 16: Accuracy of Type II Companies in Accounting

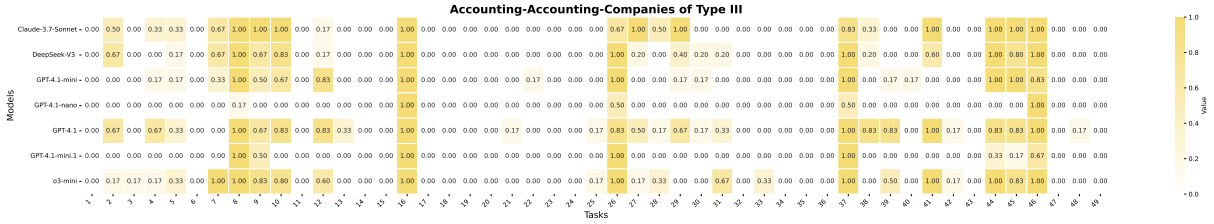


Figure 17: Accuracy of Type III Companies in Accounting

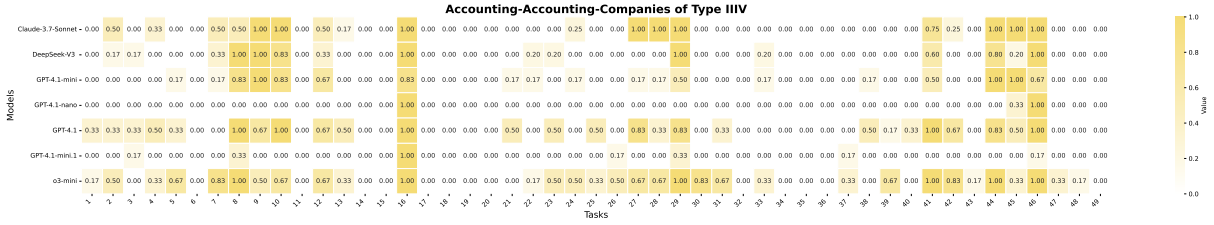


Figure 18: Accuracy of Type IIIV Companies in Accounting

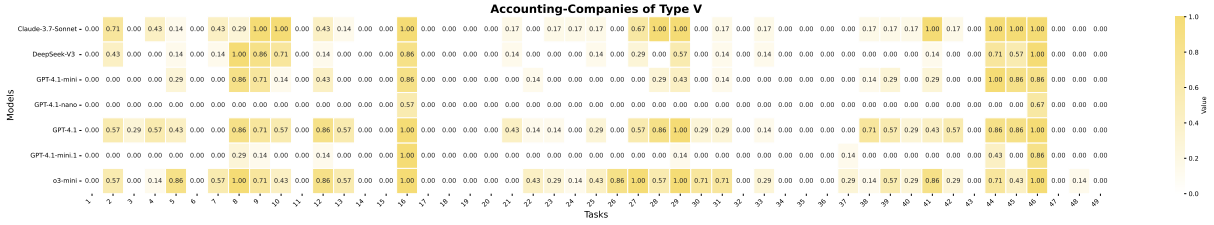


Figure 19: Accuracy of Type V Companies in Accounting

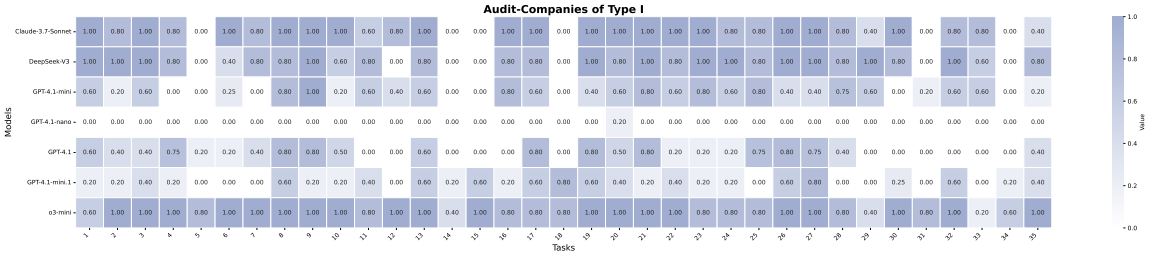


Figure 20: Accuracy of Type I Companies in Audit

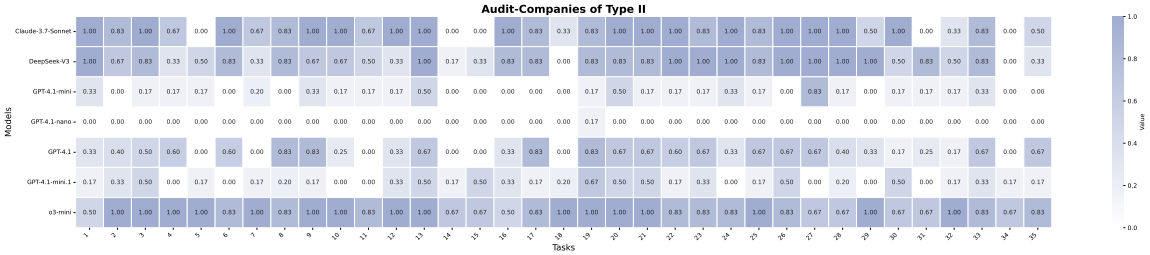


Figure 21: Accuracy of Type II Companies in Audit

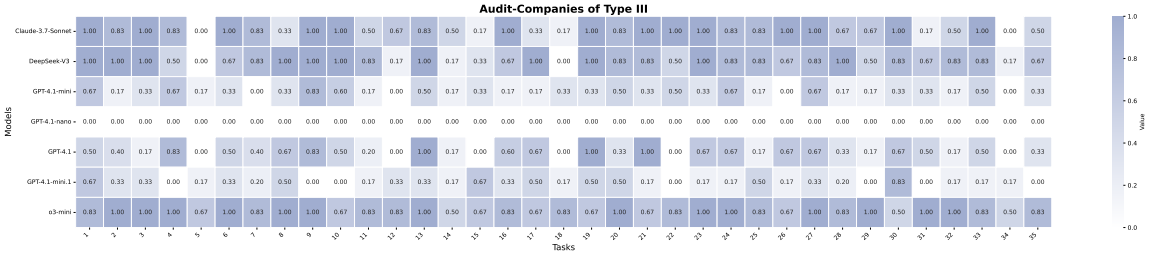


Figure 22: Accuracy of Type III Companies in Audit

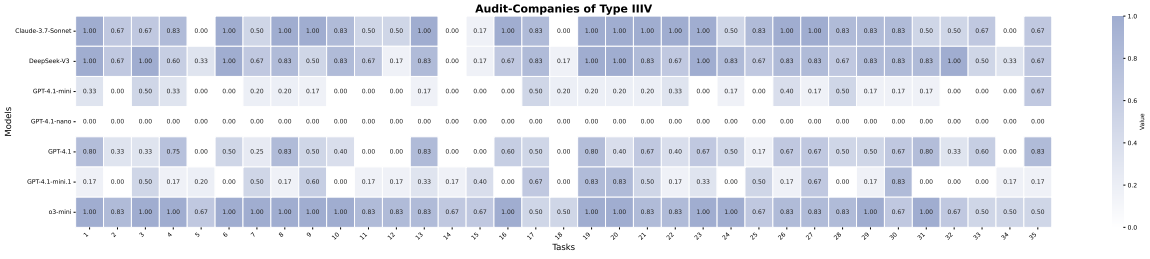


Figure 23: Accuracy of Type IIIV Companies in Audit

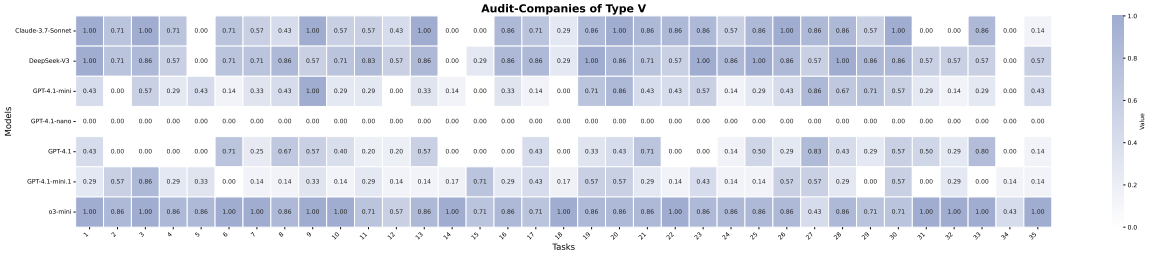


Figure 24: Accuracy of Type V Companies in Audit

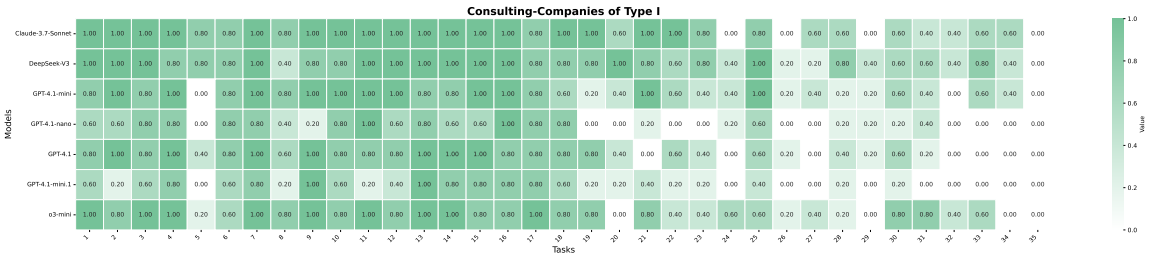


Figure 25: Accuracy of Type I Companies in Consulting

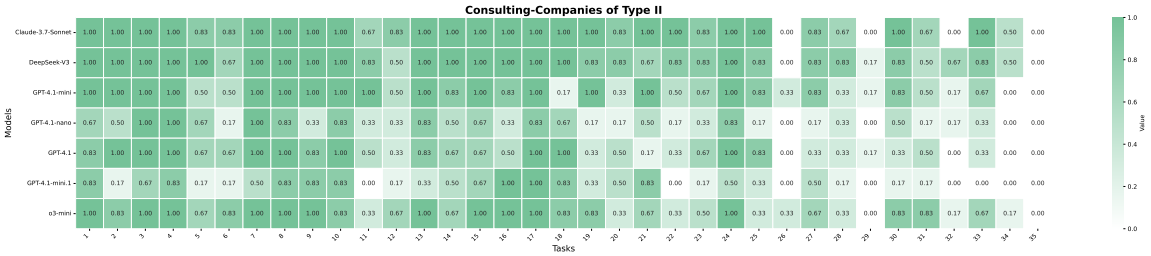


Figure 26: Accuracy of Type II Companies in Consulting

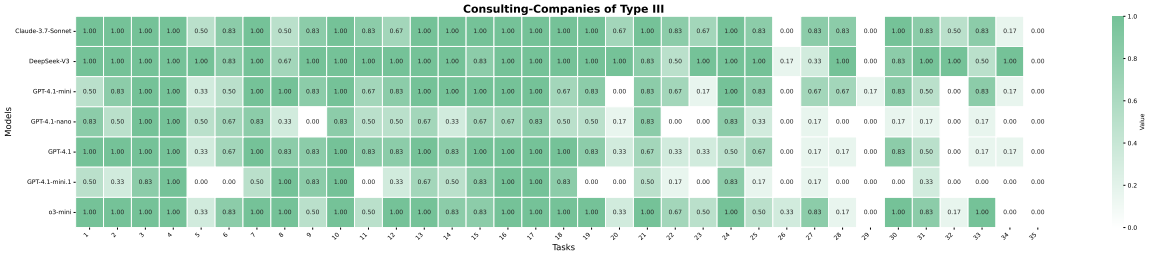


Figure 27: Accuracy of Type III Companies in Consulting

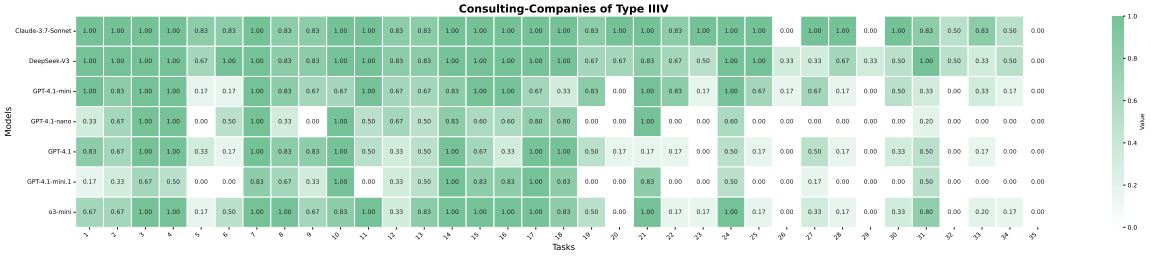


Figure 28: Accuracy of Type IIIV Companies in Consulting

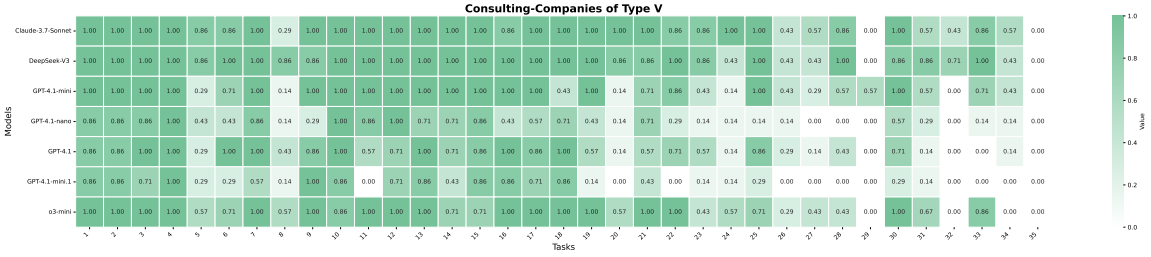


Figure 29: Accuracy of Type V Companies in Consulting

K Analysis of Reasoning Failures

To better understand why models fail, we summarize representative critical errors observed across different LLMs. We group failures into recurring error types such as factual deviation, knowledge retrieval deficiency, arithmetic and floating-point errors, format-handling issues, contextual inconsistency, critical data omission, and methodology chain breaking, and provide concrete examples for each model. This analysis shows that failures stem not only from numerical mistakes, but also from missing key facts, misinterpreting accounting and audit concepts, breaking multi-step workflows, and failing to reliably parse complex transaction inputs.

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 39: Comparative Analysis of Critical Errors Across Large Language Models

L 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 40.

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 40: Cost for LLMs