BizCompass: Benchmarking the Reasoning Capabilities of LLMs in Business Knowledge and Applications

Jianing Hao[1]*, Yuhe Wu[1]*, Yuanjian Xu[1]*, Shichang Meng[2]†, Shuai Yuan[3]†, Wei Zeng[1], Zixuan Wang[1], Guang Zhang[1]‡

[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[2]City University of Hong Kong, Hong Kong, China
[3]Peking University, Beijing, China

Abstract

Large language models (LLMs) hold great promise for business applications, yet business analysis remains inherently complex, demanding rigorous reasoning and the integration of diverse knowledge sources. Existing benchmarks typically target narrow tasks and thus leave a fundamental question unanswered: *how can LLMs be reliably applied in business, and how are these applications grounded in underlying theoretical capabilities?* To address this gap, we introduce **BizCompass**, a benchmark explicitly designed to connect theoretical foundations with practical business knowledge and applications. At the knowledge level, BizCompass covers four core domains—finance, economics, statistics, and operations management. At the application level, it structures tasks around three representative roles: the *analyst*, the *trader*, and the *consultant*. This dual-axis design not only exposes performance differences across realistic scenarios but also diagnoses which foundational capabilities enable or constrain success. We systematically evaluate both open-source and commercial LLMs, revealing how theoretical knowledge translates into practical performance in business. The results provide actionable insights for model selection and training optimization in real-world business contexts. All datasets and evaluation code are publicly released to support reproducibility and future research: https://bizcompass.dev.ypemc.com.

1 Introduction

Large language models (LLMs) have rapidly become a focal point of AI research (Brown et al., 2020; OpenAI et al., 2024; DeepSeek-AI et al., 2025), and their potential to support business analysis and decision-making is increasingly regarded as a critical direction for exploration (Zhou et al., 2024; Chen and Chan, 2024; de Kok, 2025; Siano, 2025). From financial statement analysis and portfolio optimization to strategic consulting and market forecasting, both academia and industry expect LLMs to serve as a key driver of intelligent enterprise transformation. However, business analysis itself is inherently complex: it requires not only broad interdisciplinary knowledge but also rigorous reasoning and decision-making in uncertain and dynamic environments (Eisenhardt and Zbaracki, 1992; Vahed, 2025). For example, when advising a client on entering a new market, an analyst must combine financial assessment, economic analysis, statistical evidence from surveys, and operations management methods to ensure both strategic soundness and operational feasibility. This complexity distinguishes business reasoning from general reasoning: it requires integrating multi-disciplinary knowledge within strict domain-specific constraints. For example, trading requires reconciling statistical data with economic rules, while consulting involves tracking how operational changes propagate through a firm. Existing benchmarks have made significant progress; however, they still fail to address a fundamental question for reliable LLM deployment in business:

> ❓ *Which business scenarios can LLMs reliably support, and what foundational capabilities are required to enable their effective deployment?*

Despite growing interest in applying LLMs to business decision-making, existing benchmarks remain unable to answer the fundamental question of what LLMs can reliably do in real business contexts and which foundational capabilities support such applications. While these benchmarks represent meaningful progress beyond generic NLP evaluations, they suffer from two critical limitations. First,

* Equal contribution. † Work done while the author was a Research Assistant at HKUST-GZ. ‡ Correspondence: guangzhang@hkust-gz.edu.cn

Findings of the Association for Computational Linguistics: ACL 2026, pages 23927–23966

Benchmark	Knowledge Coverage				Application Dimensions		
	Finance	Economics	Statistics	OM	Analysis	Trading	Consulting
FinQA (Chen et al., 2021a)	✓	✗	✗	✗	✓	✗	✗
CovFinQA (Chen et al., 2022)	✓	✗	✗	✗	✓	✗	✗
FinanceBench (Islam et al., 2023)	✓	✗	✗	✗	✓	✗	✗
BizBench (Krumdick et al., 2024)	✓	✓	✗	✗	✓	✗	✗
FinBen (Xie et al., 2024)	✓	✓	✓	✗	✓	✓	✗
FinMaster (Jiang et al., 2025)	✓	✓	✓	✓	✓	✗	✓
BizCompass (ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: A sample comparison between BizCompass and existing benchmarks. Red ✓ indicates full coverage and green ✗ indicates lack of coverage. A detailed comparison is shown in Appendix A.

they provide only a narrow view of real business scenarios. For example, FinQA (Chen et al., 2021a) focuses on multi-step numerical reasoning over financial reports, but its scope is largely limited to tabular calculations and does not reflect the diverse tasks that financial institutions prioritize. While FinanceBench (Islam et al., 2023) moves a step closer to practice by evaluating open-domain QA over SEC filings, its problem formulation remains narrow. It primarily tests single-turn, isolated questions, which falls short of the complex information synthesis required in real-world professional workflows. Second, current benchmarks lack a systematic bridge between theoretical capabilities and practical applications. Some remain overly abstract and detached from real-world practice, while others emphasize narrow applications without clarifying the foundational skills required. For instance, although FinanceQA (Mateega et al., 2025) simulates professional investment workflows, it does not evaluate the underlying foundational knowledge. Consequently, when mainstream models fail on nearly 60% of cases, it is impossible to diagnose which specific capability gaps caused the failure, underscoring the absence of a clear mapping from theoretical knowledge to applied performance. As a result, prior efforts offer fragmented insights rather than a coherent framework, leaving the connection between disciplinary foundations and real business applications unresolved. An abstract comparison is shown in Table 1 to demonstrate our advantages. The detailed comparison between BizCompass and other past 5-year representative benchmarks is provided in Table 5 in Appendix A.

To fill this critical gap and enable the reliable deployment of LLMs in business, we introduce **BizCompass**, the first benchmark that systematically bridges disciplinary foundations with realistic business applications. BizCompass offers an integrated evaluation framework tailored to high-stakes business contexts. Its design is guided by two central principles. First, rigorous evaluation must be grounded in *realistic business scenarios*: to this end, we engaged in extensive discussions with over ten leading financial institutions to identify the application settings they regard as most critical. Second, meaningful assessment requires *linking these scenarios back to foundational capabilities*: we therefore constructed tasks that capture both disciplinary knowledge and its applied manifestations. Our contributions go substantially beyond prior efforts and can be summarized as follows:

- **Systematic coverage:** BizCompass is the first benchmark to integrate four foundational disciplines—finance, economics, statistics, and operations management—together with role-aware tasks spanning analysis, trading, and consulting. This structured design ensures that both theoretical underpinnings and practical applications are comprehensively represented.
- **Explanatory evaluation:** BizCompass moves beyond surface-level task accuracy by explicitly linking performance to underlying theoretical capabilities. This enables not only the measurement of task success or failure, but also the diagnosis of which foundational skills drive or hinder performance in real-world business contexts.
- **Reliable and rigorous data:** All tasks were co-developed with domain experts and subjected to a systematic, multi-stage review process. Coupled with a substantial volume of carefully designed questions, this guarantees realism, statistical robustness, and minimal evaluation bias, thereby establishing a new standard for business-oriented LLM benchmarks.

Through this **carefully constructed framework**,

BizCompass establishes the first systematic mapping between foundational capabilities and applied business scenarios. In doing so, it uncovers critical performance gaps of current LLMs, while providing actionable guidance for model selection, training strategies, and real-world deployment.

2 Benchmark Construction

In evaluating domain knowledge for business applications, we focus on Finance, Economics, Statistics, and Operations Management, which jointly provide a coherent analytical framework. Their relevance is supported by both the curricular structures of leading business schools and prior research (Biehl et al., 2006). To bridge this disciplinary foundation and practical applications, we draw on the O*NET occupational classification system[1] to identify relevant professional categories. We find that analysis, trading, and consulting collectively span over 80% of business scenarios. Further details are provided in the Appendix J.1.

2.1 Data Sources

To bridge theory and practice, we curate a high-quality corpus to construct BizCompass and organize it into two categories: one for constructing knowledge-based tasks and the other for designing application-based scenarios.

Data Sources for Knowledge-Based Tasks. For knowledge-based tasks, we rely on academic journal articles, motivated by two main considerations. First, their peer-reviewed nature makes them reliable sources of knowledge. Second, they combine methodological rigor with both established and emerging theories, providing a solid foundation for testing whether LLMs can function as domain experts. We select 82 representative journals across economics, finance, operations management, and statistics, referencing recommended lists from top universities and prior research (Ham et al., 2021). The corpus spans 1954–2025, thus covering both classic theories—often only briefly covered in textbooks—and frontier research. In terms of quantity, it comprises 152,077 articles, sufficient to ensure a comprehensive and unbiased evaluation. Further details are provided in Appendix B.

Data Sources for Application-Based Tasks. Before developing application-based tasks, we systematically reviewed critical business scenarios and existing benchmarks. This analysis shows that current benchmarks remain limited, with coverage restricted to a few tasks, exemplified by CCFraud (Feng et al., 2023). To address the lack of benchmarks for other tasks, we established three principles for material selection. Authenticity requires materials to mirror real-world contexts, in line with financial analysis practices of grounding conclusions in public documents (Loughran and McDonald, 2011). Utility requires materials to include step-by-step operational details, reflecting the practical need in financial machine learning to turn theory into deployable strategies (De Prado, 2018). Governance requires clear source traceability and usage rights, in line with standards such as Datasheets for Datasets (Gebru et al., 2021) and Model Cards (Mitchell et al., 2019). Based on the above principles, we primarily rely on two types of source materials:

- **Practitioner-Oriented Textbooks:** Practitioner-oriented textbooks serve as our primary source for practical knowledge (De Prado, 2018; Chan, 2013). We convert their stepwise workflows, methods, and constraints into multi-step application tasks. For instance, the chapter of the Sharpe ratio—including its definition, formula, and interpretation—is converted into a composite task where a model must calculate the ratio from given asset returns, analyze its implications for risk-adjusted performance, and provide a final investment recommendation.
- **Real-World Business Documents:** Real-world business documents, in turn, provide the raw business context and factual grounding for our evaluation tasks (Loughran and McDonald, 2011). These materials include corporate disclosures (10-Ks, 10-Qs)[2] for fundamental analysis, third-party structured datasets (Morningstar)[3] for quantitative reasoning, and newspaper archives for event-driven analysis.

2.2 Construction Pipeline

We design a three-phase pipeline comprising data preprocessing, question construction, and evalua-

[1] National Center for O*NET Development. ONET-SOC Taxonomy. https://www.onetcenter.org/taxonomy.html

[2] The 10-K is an annual report and the 10-Q is a quarterly report filed by public companies with the U.S. Securities and Exchange Commission (SEC). Available via the SEC's EDGAR database: https://www.sec.gov/edgar/searchedgar/companysearch

[3] The morningstar is a global financial services firm that provides extensive financial data and investment research on a wide range of securities: https://www.morningstar.com/

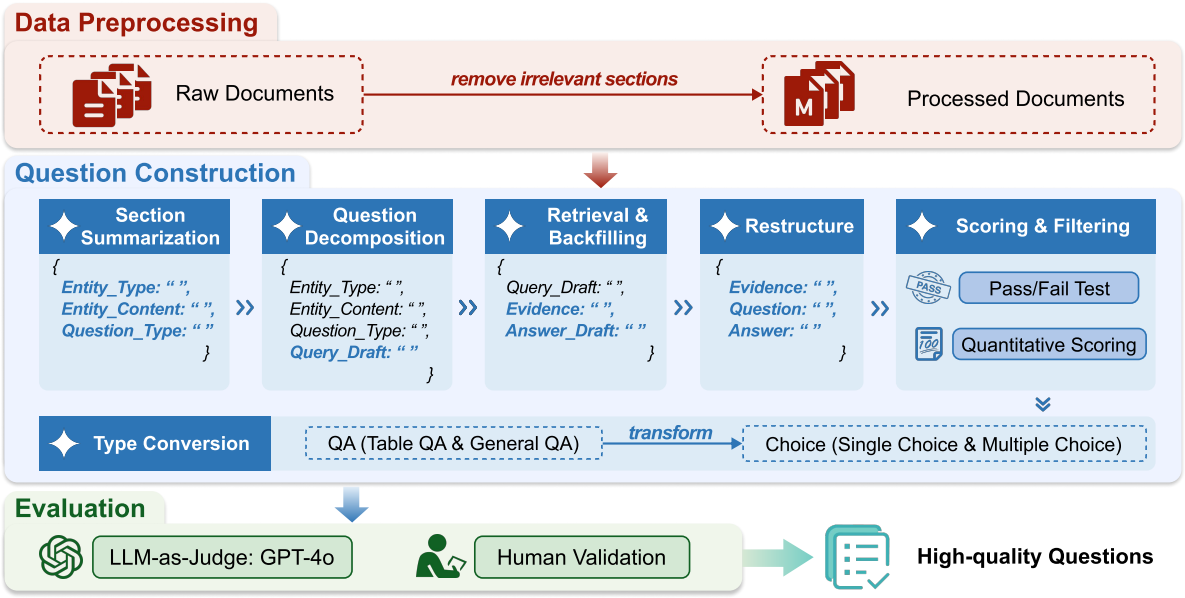

Figure 1: The three-phase pipeline of BizCompass benchmark construction.

tion, as illustrated in Figure 1. In the preprocessing phase, raw PDFs are converted and cleaned into an analysis-ready corpus. The construction phase follows six modular steps: (i) section summarization extracts key knowledge units such as equations, propositions, and tables; (ii) question decomposition generates candidate queries; (iii) retrieval and backfilling ground each query in verbatim textual evidence; (iv) restructuring organizes evidence and queries into coherent question–answer pairs; (v) scoring and filtering ensure consistency and quality; and (vi) type conversion adapts each item into QA or choice format. This stepwise design makes the process auditable and reduces error propagation. The final evaluation phase adopts a dual-track expert review, where specialists assess knowledge-based tasks and practitioners assess application-based ones. These iterative reviews focused on clarity, representativeness, realism, and difficulty. We adopted a *"Review-Discuss-Resolve"* protocol to handle disagreements. Convergence was defined as achieving unanimous consent; if consensus could not be reached after two rounds of discussion, the item was strictly discarded. Detailed descriptions of each step are provided in Appendix C.

2.3 BizCompass Statistics

BizCompass contains a total of 14,855 questions. Among them, 6,406 are knowledge-based questions, and 8,449 are application-based questions. The distribution of the data is illustrated in Figure 2. The inner circle of the figure represents four categories of knowledge-based questions and three major application domains, namely Consulting, Trading, and Analysis. The middle and outer layers further expand into specific sub-tasks and question formats, such as question answering, single choice, and multiple choice.

The design of BizCompass carefully considers the varying contextual length requirements across different task types, as shown in Figure 3. Some tasks naturally require long-context analytical capabilities, involving cross-sentence reasoning, information extraction, or document-level comprehension. In contrast, other interactive scenarios rely on shorter contexts with more frequent exchanges rather than lengthy narratives. This balanced design ensures that BizCompass aligns with real-world business communication patterns, thereby enabling more accurate evaluation of LLMs.

3 Experiments

BizCompass is designed to support model selection for business applications and to provide guidance on how LLMs can be further optimized. To meet these objectives, we structure our experiments around three research questions that establish performance baselines, explain underlying causes, and identify pathways for improvement:

- **RQ1 (What):** What is the overall performance of LLMs on BizCompass, and what is the relationship between application scenarios and knowledge domains in shaping performance?

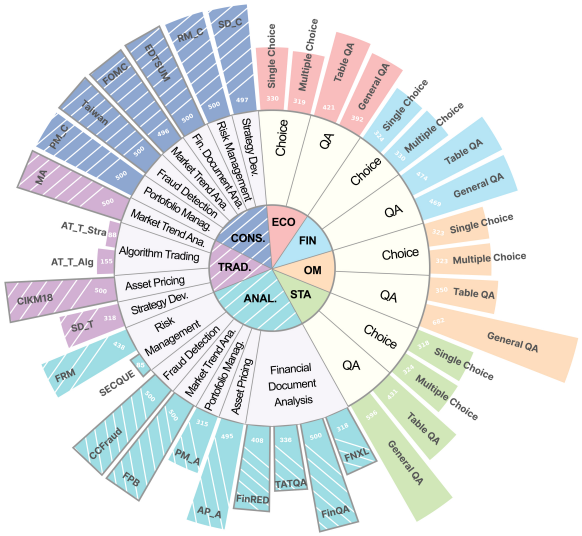

Figure 2: BizCompass's statistics. Bordered bars indicate subsets derived from existing datasets, while non-bordered bars represent newly created original subsets by BizCompass. For consistency, we define the abbreviations as follows: **AP** = Asset Pricing, **MTA** = Market Trend Analysis, **RM** = Risk Management, **SD** = Strategy Development, **PM** = Portfolio Management, **AT** = Algorithmic Trading (with suffixes **Stra** = Strategy and **Alg** = Algorithm), **FD** = Fraud Detection, and **FDA** = Financial Document Analysis. **A**, **T**, and **C** denote the three major application domains: Analysis, Trading, and Consulting, respectively.

- **RQ2 (Why):** What factors underlie performance differences, and how are these reflected in the reasoning paths of models when solving complex business problems?

- **RQ3 (How):** How can appropriate fine-tuning methods be employed to enhance the performance of LLMs in business scenarios?

To provide a solid foundation for investigating these research questions, we begin by outlining the experimental setup, focusing on model selection and fairness control.

3.1 Evaluation Setup

Model Selection. We systematically evaluate 23 representative LLMs under the few-shot setting. To ensure comprehensive coverage, the evaluated LLM series includes both open-source and proprietary models, such as GPT (OpenAI, 2024), Gemini (Comanici et al., 2025), Llama (AIMeta, 2025), Claude (Anthropic, 2025), DeepSeek (DeepSeek-AI et al., 2025), Qwen (Team, 2024), and Grok (xAI, 2025). Our evaluation also includes distilled models (*e.g.*, 'DeepSeek-R1-Distill-Llama-70B' (Tian et al., 2025)), and models with integrated thinking mechanisms for efficient comparisons. To facilitate a fair and standardized comparison, we consistently employ the official default configurations for models with adjustable reasoning capacities.

Evaluation Metrics. We employ a diverse set of evaluation metrics tailored to the specific nature of each task.

- **Accuracy.** For single-choice questions and tasks with label-based outputs, we use Accuracy to measure the proportion of correct predictions, which provides a straightforward assessment of model performance on classification tasks.

- **F1.** For multi-choice questions, we complement Accuracy with F1 score to better capture both precision and recall, providing a more balanced evaluation of model performance on multi-label classification tasks.

- **ROUGE.** For tasks requiring text generation, we employ ROUGE-1, ROUGE-2, and ROUGE-L metrics and compute their average to provide a comprehensive assessment of text similarity (Lin, 2004). This averaging approach ensures a more balanced evaluation across different levels of n-gram matching and longest common subsequence matching.

- **GPT-Eval.** Considering the dynamic nature of answers and the need for understanding the logic of answers, we adopt LLM-as-a-Judge (Zheng et al., 2023) for evaluation, name the metric GPT-Eval and select GPT-4o as our evaluator. To actively mitigate potential evaluation biases, we design structured, dimension-specific scoring criteria inspired by G-Eval (Liu et al., 2023) (The prompts are detailed in Appendix I). Recent studies, such as JudgeBench (Tan et al., 2025), have validated GPT-4o's strong alignment with human judges, thereby justifying its selection as a reliable evaluator.

Fairness control. To ensure fair and comparable evaluations across diverse models, we standardize the hyperparameters via a comprehensive grid search on the Llama3-8B model, exploring temperature values in {0, 0.2, 0.4, 0.6, 0.8} and top-p values in {0.6, 0.8, 0.9, 0.95}. For multiple-choice questions in knowledge-based tasks, accuracy is maximized at a temperature of 0.8 and a top-p of

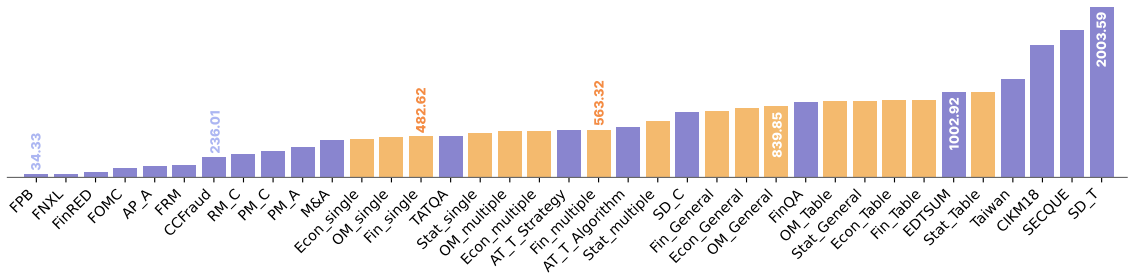

Figure 3: The average token length of each subset in BizCompass.

0.95. For general QA benchmarks, where GPT-4o is employed as the evaluator following the protocol in Appendix I, the optimal configuration shifts to a temperature of 0.8 paired with a top-p of 0.6. Detailed outcomes of the grid search are provided in Appendix F, specifically in Table 13.

3.2 Evaluation Results

As shown in Tables 2 and 3, proprietary LLMs (e.g., GPT, Gemini, Claude) outperform open-source models in both domain knowledge and practical application tasks. In knowledge-based evaluation, proprietary models show higher consistency across finance, economics, statistics, and operations management, suggesting broader coverage of professional corpora. In contrast, the performance gap widens in application-based tasks that require contextual reasoning and multi-step decision making. Larger model scales do not necessarily lead to higher accuracy, as DeepSeek-R1 performs worse than some smaller proprietary LLMs, indicating that scaling alone is not sufficient. Likewise, incorporating a chain-of-thought mechanism (Wei et al., 2022a) does not guarantee improvement, implying that reasoning effectiveness depends more on data quality and alignment than on explicit reasoning traces. We can conclude that domain knowledge, contextual reasoning, model scale, and thinking mechanisms contribute differently to business performance, and their effects are not additive.

4 Discussion

Our results highlight a relationship between knowledge, reasoning, and application performance. To better understand these findings, our discussion first addresses RQ2 to diagnose the factors driving performance differences. Building on this diagnostic insight, we then explore potential remedies by addressing RQ3.

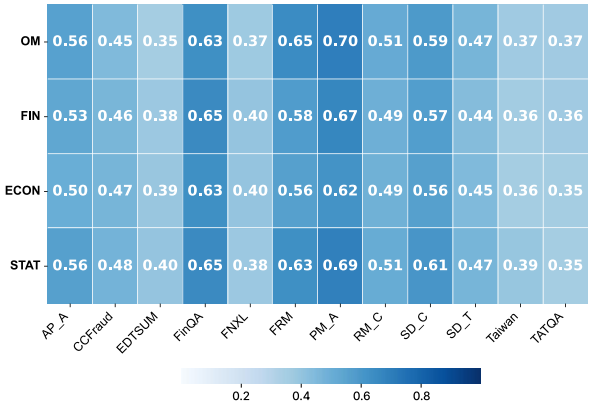

Figure 4: Correlation between application-based tasks and the 4 knowledge domains.

4.1 Correlation Analysis

Cross-domain Correlation. As shown in Figure 4, most tasks are influenced by domain differences, indicating that domain representations systematically shape task features. Some tasks show stronger domain dependence: FRM and PM_A are more correlated with OM and STAT, suggesting shared structural patterns between statistical modeling and operational decision-making. In contrast, FIN and ECON exhibit weaker and more uniform correlations, implying looser structural connections. From the task perspective, analytical and quantitative tasks display higher cross-domain consistency, while text and consulting tasks show lower correlations, reflecting their stronger reliance on semantic and contextual information.

Related to Code Reasoning Ability. We evaluate the relationship between models' ability to solve complex code reasoning problems and their performance on the BizCompass. As shown in Figure 5 (A), model performance on SWE-bench (Jimenez et al., 2024) is positively correlated with knowledge-based tasks in BizCompass. This indicates that the cognitive skills involved in complex code reasoning, such as decompositional

Model	Size	Thinking	FIN		ECON		OM		STAT	
			Acc.	GPT-Eval	Acc.	GPT-Eval	Acc.	GPT-Eval	Acc.	GPT-Eval
Proprietary LLMs										
GPT	N/D	✓	80.43%	4.98	83.03%	4.98	79.26%	4.98	83.80%	4.99
Gemini	N/D	✓	82.13%	4.95	87.77%	4.95	82.66%	4.93	85.67%	4.97
Claude	N/D	✓	81.82%	4.90	85.78%	4.92	80.18%	4.89	84.58%	4.93
Grok	N/D	✓	76.58%	4.94	83.03%	4.93	76.63%	4.90	78.51%	4.96
Open-source LLMs										
DeepSeek	671B	✓	73.81%	4.91	81.65%	4.91	71.05%	4.89	70.87%	4.95
Llama	70B	✗	52.59%	4.66	62.84%	4.74	60.52%	4.61	57.79%	4.71
Qwen	235B	✓	78.58%	4.86	81.65%	4.87	80.03%	4.87	82.09%	4.91
Distilled LLMs										
DeepSeek-R1-Distill	70B	✓	70.11%	4.56	79.97%	4.68	72.13%	4.54	71.96%	4.70

Table 2: Performance of different models on knowledge-based tasks of BizCompass. For each model series, we report the model with highest average accuracy and GPT-Eval scores. N/D means the parameter size is not disclosed. The complete results of all selected LLMs are presented in Table 15 in Appendix H.

Model	Size	Thinking	8-task Avg.	3-task Avg.	4-task Avg.	6-task Avg.
			Accuracy	ROUGE	GPT-Eval (Label)	GPT-Eval (Score)
Proprietary LLMs						
GPT	N/D	✓	79.94%	24.30%	52.35%	4.82
Claude	N/D	✓	75.50%	30.73%	56.86%	4.43
Gemini	N/D	✓	77.43%	23.68%	50.92%	4.47
Grok	N/D	✓	78.40%	21.09%	50.13%	4.25
Open-source LLMs						
DeepSeek	671B	✗	71.26%	19.58%	44.97%	4.31
Llama	70B	✗	60.24%	25.15%	41.93%	4.01
Qwen	235B	✓	64.78%	14.12%	49.56%	4.37
Distilled LLMs						
Deepseek-R1-Distill	32B	✓	62.30%	24.32%	45.06%	4.11

Table 3: Performance of different models on application-based tasks of BizCompass. For each model series, we report the one with the best performance. It is computed as the weighted average of metrics across tasks using the same metric. The complete results of all selected LLMs are presented in Table 16 in Appendix H.

thinking, consistency checking, and structured information manipulation, also strengthen a model's capacity for factual recall and conceptual integration. For application-oriented tasks, the correlation remains positive but becomes less pronounced at higher performance levels. Once a model's SWE-bench accuracy surpasses approximately 60%, further improvements in code reasoning contribute little to its downstream application performance. This suggests that real-world application ability relies not only on symbolic reasoning but also on contextual understanding and the transfer of reasoning to business settings.

Relation to Long-Context Ability. We further analyze how long context understanding relates to model performance on BizCompass. As shown in Figure 5 (B), performance on LongBench v2 (Bai et al., 2024) is positively correlated with both knowledge-based and application-based tasks in BizCompass. Models with stronger long-context abilities tend to achieve higher overall scores, but the relationship varies across task types. For knowledge-based tasks, the correlation is stronger yet segmented across models, while application-based tasks display a smoother and more consistent trend. Models such as Gemini-2.5-Pro and Qwen-3 effectively leverage extended context for cross-segment reasoning and knowledge integration, whereas others with similar LongBench accuracy, such as GPT-4o and DeepSeek-R1, show smaller gains. This suggests that long context capacity alone does not guarantee better knowledge reasoning; performance depends on how well a model utilizes extended context to preserve coherence and perform higher-level inference. In contrast, application tasks rely more on reasoning-

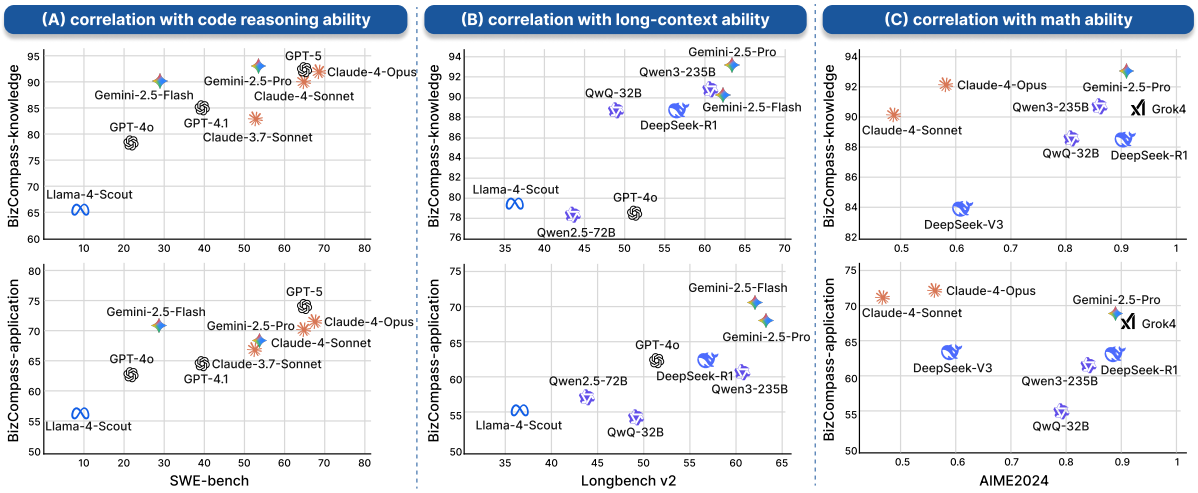

Figure 5: Correlation of model performance on (A) SWE-bench (resolved %), (B) LongBench v2 (accuracy %), and (C) AIME2024 (10× accuracy rate %) with weighted scores from BizCompass knowledge-based and application-based evaluations.

to-action transfer, where longer context provides modest but more uniform benefits.

Relation to Math Ability. Figure 5 (C) shows the relationship between model performance on BizCompass-application tasks and their AIME2024 math scores. The correlation appears weak. Models with stronger mathematical reasoning, such as Gemini-2.5-Pro and DeepSeek-R1, achieve higher scores on business-oriented tasks, but their advantages are limited. While basic mathematical competence provides a foundation for structured reasoning, further improvements in business performance depend on broader capabilities such as combining diverse information, identifying implicit objectives, and handling incomplete or changing contexts. These abilities rely more on pragmatic understanding and adaptive alignment than on numerical skills. Therefore, mathematical and business reasoning represent distinct yet partially overlapping cognitive capacities.

Question types	Model		
	Llama3-8B	SFT_1[4]	SFT_2[5]
SC. (%)	38.89	51.23 (+12.34)	37.96 (-0.93)
MC. (%)	14.55	26.97 (+12.42)	11.21 (-3.34)
TableQA	3.24	3.28 (+0.04)	3.11 (-0.13)
GeneralQA	2.77	2.79 (+0.02)	2.56 (-0.21)

Table 4: Performance comparison of 2 SFT models against the Llama3-8B baseline across different question types in Finance domain. Red indicates improvement, and green indicates decrease. SC. and MC. are evaluated by accuracy (%), while TableQA and GeneralQA are evaluated by overall scores.

4.2 Domain-specific SFT

In this section, we address RQ3, which investigates whether domain-specific fine-tuning can effectively enhance the reasoning ability of LLMs, using financial tasks as a representative case. To ensure our evaluation reflects practical deployment constraints, inference experiments for the fine-tuned models were conducted on 4×NVIDIA A40 GPUs. As shown in Table 4, targeted knowledge injection through supervised fine-tuning improves model performance on simple knowledge-based financial questions. The SFT model trained on the Sujet-Finance-Instruct dataset achieves substantial gains in single-choice and multi-choice tasks, indicating that domain-specific supervision effectively enhances factual recall and pattern recognition. However, its improvement on more complex tasks such as TableQA and GeneralQA is marginal, suggesting that fine-tuning primarily benefits surface-level knowledge acquisition rather than deeper reasoning. These findings imply that while domain instruction tuning strengthens localized knowledge representation, it contributes little to the model's ability to perform compositional or context-dependent reasoning. Furthermore, to provide practical guidance for real-world deployment, we systematically profiled the computational overhead of the fine-tuned model; detailed performance metrics are provided in Appendix G.

[4] The model is fine-tuned on the sujet-ai/Sujet-Finance-Instruct-177k dataset.

[5] The model is fine-tuned on the Josephgflowers/Finance-Instruct-500k dataset.

5 Conclusion

We present BizCompass, a benchmark that unifies domain knowledge and real-world business reasoning to systematically evaluate LLMs. Extensive experiments reveal that while proprietary models outperform open-source ones, both suffer from reasoning bottlenecks in compositional, multi-step, and cross-domain tasks. Domain-specific fine-tuning improves factual accuracy but offers limited gains in contextual reasoning, highlighting the need to integrate structured knowledge with adaptive reasoning alignment. BizCompass establishes a foundation for future research on model diagnosis, reasoning enhancement, and domain-adaptive alignment in business-critical contexts.

6 Limitations

BizCompass currently focuses on text-based reasoning. Future work will extend it toward multimodal business understanding, as many real-world tasks require interpreting images, charts, and other visual materials alongside text. Incorporating these modalities, together with interactive and decision-oriented evaluations, will enable a more comprehensive assessment of business intelligence in large language models.

References

AIMeta. 2025. The llama4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed:2025-04-05.

Lorin W Anderson and David R Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Addison Wesley Longman, Inc.

Anthropic. 2025. Claude3.7 sonnet and claude code. Accessed:2025-02-25.

American Statistical Association and 1 others. 2024. Overview of Statistics as a Scientific Discipline and Practical Implications for the Evaluation of Faculty Excellence.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. LongBench: A Bilingual, Multitask Benchmark for Long Context Understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Edoardo Bajo, Massimiliano Barbi, and David Hillier. 2020. Where should i publish to get promoted? a finance journal ranking based on business-school promotions. *Journal of Banking & Finance*, 114:105780.

Deborah L Bandalos. 2018. *Measurement Theory and Applications for the Social Sciences*. Guilford Publications.

Noga BenYoash, Menachem Brief, Oded Ovadia, Gil Shenderovitz, Moshik Mishaeli, Rachel Lemberg, and Eitam Sheetrit. 2025. SECQUE: A Benchmark for Evaluating Real-World Financial Analysis Capabilities. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 212–230.

Vincent Berthet. 2022. The Impact of Cognitive Biases on Professionals' Decision-Making: A Review of Four Occupational Areas. *Frontiers in psychology*, 12:802439.

Markus Biehl, Henry Kim, and Michael Wade. 2006. Relationships among the academic business disciplines: A multi-method citation analysis. *Omega*, 34(4):359–371.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. 2025. Finance Agent Benchmark: Benchmarking LLMs on Real-world Financial Research Tasks. *arXiv preprint arXiv:2508.00828*.

Kenneth A. Borokhovich, Robert J. Bricker, and Betty J. Simkins. 2000. An Analysis of Finance Journal Impact Factors. *Journal of Finance*, 55(3):1457–1469.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Egon Brunswik. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological review*, 62(3):193.

Christopher JC Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 89–96.

Saul Carliner, Chantal Castonguay, Emily Sheepy, Ofelia Ribeiro, Hiba Sabri, Chantal Saylor, and Andre Valle. 2015. The job of a performance consultant: a qualitative content analysis of job descriptions. *European Journal of Training and Development*, 39(6):458–483.

Ernest P. Chan. 2013. *Algorithmic Trading: Winning Strategies and Their Rationale*. John Wiley & Sons.

Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. FinTextQA: A Dataset for Long-form Financial Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and 1 others. 2023. DISC-FinLLM: A Chinese Financial Large Language Model based on Multiple Experts Fine-tuning. *arXiv preprint arXiv:2310.15205*.

Zenan Chen and Jason Chan. 2024. Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. *Management Science*, 70(12):9101–9117.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021a. FinQA: A Dataset of Numerical Reasoning over Financial Data. *Proceedings of EMNLP 2021*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021b. FinQA: A Dataset of Numerical Reasoning over Financial Data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 6279–6292.

Janis Cloos, Matthias Greiff, and Hannes Rusch. 2023. Editorial favoritism in the field of laboratory experimental economics. *Journal of Behavioral and Experimental Economics*, 107:102082.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Pierre-Philippe Combes and Laurent Linnemer. 2010. Inferring Missing Citations: A Quantitative Multi-Criteria Ranking of all Journals in Economics. HAL Working Paper.

Robert Currie and G. S. Pandher. 2020. Finance journal rankings: Active scholar assessment revisited. *Journal of Banking & Finance*, 111:105717.

Ties de Kok. 2025. ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science*.

Marcos Lopez De Prado. 2018. *Advances in Financial Machine Learning*. John Wiley & Sons.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025. DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638.

Yu Ding. 2023. Editorial: Perspectives of ISE/OR researchers. *IISE Transactions*, 55(1):1–1.

Forty-ninth Edition. 2000. Journal Quality List.

Kathleen M Eisenhardt and Mark J Zbaracki. 1992. Strategic decision making. *Strategic management journal*, 13(S2):17–37.

Kristie M. Engemann and Howard J. Wall. 2009. A Journal Ranking for the Ambitious Economist. *Federal Reserve Bank of St. Louis Review*, 91(3):127–139.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Alejandro Lopez-Lira, and Hao Wang. 2023. Empowering Many, Biasing a Few: Generalist Credit Scoring through Large Language Models. *arXiv preprint arXiv:2310.00566*.

Sophie Forster and Nilli Lavie. 2008. Failures to Ignore Entirely Irrelevant Distractors: The Role of Load. *Journal of Experimental Psychology: Applied*, 14(1):73.

Harry BG Ganzeboom. 2010. A standard international socio-economic index of occupational status. In *annual conference of international social survey programme, Lisbon*, volume 1.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Mark J Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. 2017. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. *Review of educational research*, 87(6):1082–1116.

Xin Guo, Haotian Xia, Zhaowei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, and 1 others. 2025. FinEval: A Chinese financial domain knowledge evaluation benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the*

Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6258–6292.

Thomas M Haladyna. 2004. *Developing and Validating Multiple-Choice Test Items*. Routledge.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied measurement in education*, 15(3):309–333.

John C Ham, Julian Wright, and Ziqiu Ye. 2021. New Rankings of Economics Journals: Documenting and Explaining the Rise of the New Society Journals. *Available at SSRN 3606030*.

Daniel S. Hamermesh. 2018. Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56(1):115–156.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, and 1 others. 2024. Retrieval-Augmented Generation with Graphs (GraphRAG). *arXiv preprint arXiv:2501.00309*.

Stephen N Haynes, David C Richard, and Edward S Kubany. 1995. Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological assessment*, 7(3):238.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering General Artificial Intelligence with Open Datasets. *arXiv preprint arXiv:2407.13773*.

James J. Heckman and Shruti Moktan. 2020. Publishing and promotion in economics: The tyranny of the Top Five. *Journal of Economic Literature*, 58(2):419–470.

Chenyu Huang, Zhengyang Tang, Shixi Hu, Ruoqing Jiang, Xin Zheng, Dongdong Ge, Benyou Wang, and Zizhuo Wang. 2025. ORLM: A Customizable Framework in Training Large Models for Automated Optimization Modeling. *Operations Research*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2391–2401.

Robert Hudson. 2025. How does the UK Academic Journal Guide compare to other journal rating guides? *Scientometrics*, pages 1–35.

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*.

Junzhe Jiang, Chang Yang, Aixin Cui, Sihan Jin, Ruiyu Wang, Bo Li, Xiao Huang, Dongning Sun, and Xinrun Wang. 2025. FinMaster: A Holistic Benchmark for Mastering Full-Pipeline Financial Workflows with LLMs. *arXiv preprint arXiv:2505.13533*.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *The Twelfth International Conference on Learning Representations*.

Pantelis Kalaitzidakis, Theofanis P. Mamuneas, and Theodore Stengos. 2011. An updated ranking of academic journals in economics. *Canadian Journal of Economics*, 44(4):1525–1538.

Ron Kaniel and Hong Liu. 2006. So What Orders Do Informed Traders Use? *The Journal of Business*, 79(4):1867–1913.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 252–262. Association for Computational Linguistics.

Yolanda K. Kodrzycki and Pingkang Yu. 2006. New Approaches to Ranking Economics Journals. *The B.E. Journal of Economic Analysis & Policy*, 5(1):000101515153806451520.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A Quantitative Reasoning Benchmark for Business and Finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332.

Suzanne Lane, Mark R Raymond, Thomas M Haladyna, and 1 others. 2016. *Handbook of Test Development*, volume 2. Routledge New York, NY.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in neural information processing systems*, volume 33, pages 9459–9474.

Shuyue S Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang W Koh, and Yulia Tsvetkov. 2024a. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024b. AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 773–783.

Chenwei Lin, Hanjia Lyu, Xian Xu, and Jiebo Luo. 2025. INS-MMBench: A Comprehensive Benchmark for Evaluating LVLMs' Performance in Insurance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9047.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

Shu Liu, Shangqing Zhao, Chenghao Jia, Xinlin Zhuang, Zhaoguang Long, Jie Zhou, Aimin Zhou, Man Lan, and Yang Chong. 2025. FinDABench: Benchmarking Financial Data Analysis Ability of Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 710–725.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Tim Loughran and Bill McDonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. BBT-Fin: Comprehensive Construction of Chinese Financial Domain Pre-trained Language Model, Corpus and Benchmark. *arXiv preprint arXiv:2302.09432*.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Spencer Mateega, Carlos Georgescu, and Danny Tang. 2025. FinanceQA: A Benchmark for Evaluating Financial Analysis Capabilities of Large Language Models. *arXiv preprint arXiv:2501.18062*.

Samuel Messick. 1995. VALIDITY OF PSYCHOLOGICAL ASSESSMENT: VALIDATION OF INFERENCES FROM PERSONS' RESPONSES AND PERFORMANCES AS SCIENTIFIC INQUIRY INTO SCORE MEANING. *American psychologist*, 50(9):741.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Raymond S Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of general psychology*, 2(2):175–220.

Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, and 1 others. 2025. CFinBench: A comprehensive Chinese financial benchmark for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 876–891.

OpenAI. 2024. Hello gpt-4o. Accessed:2024-05-13.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report.

Yasuhiro Ozuru, Stephen Briner, Christopher A Kurby, and Danielle S McNamara. 2013. Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(3):215.

Hendrik Radatz. 1979. Error analysis in mathematics education. *Journal for Research in mathematics Education*, 10(3):163–172.

Javier Ruiz-Castillo and Ludo Waltman. 2015. Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, 9(1):102–117.

Bonnie R Rush, David C Rankin, and Brad J White. 2016. The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC medical education*, 16(1):250.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. In *The Twelfth International Conference on Learning Representations*.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. In *Proceedings of the 61st*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6664–6679. Association for Computational Linguistics.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

James Shanteau. 1992. Competence in experts: The role of task characteristics. *Organizational behavior and human decision processes*, 53(2):252–266.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial Numeric Extreme Labelling: A Dataset and Benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561.

Soumya Sharma, Tapas Nayak, Arusarka Bose, Ajay Kumar Meena, Koustuv Dasgupta, Niloy Ganguly, and Pawan Goyal. 2022. FinRED: A Dataset for Relation Extraction in Financial Domain. In *Companion Proceedings of the Web Conference 2022*, pages 595–597.

Federico Siano. 2025. The News in Earnings Announcement Disclosures: Capturing Word Context Using LLM Methods. *Management Science*.

Steven A Sloman. 1993. Feature-based induction. *Cognitive psychology*, 25(2):231–280.

Xinyi Song, Lina Lee, Kexin Xie, Xueying Liu, Xinwei Deng, and Yili Hong. 2026. StatLLM: A Dataset for Evaluating the Performance of Large Language Models in Statistical Analysis. *Scientific Data*.

Stephen M Stigler. 1994. Citation Patterns in the Journals of Statistics and Probability. *Statistical Science*, pages 94–108.

Samuel J Stratton. 2024. Purposeful Sampling: Advantages and Pitfalls. *Prehospital and disaster medicine*, 39(2):121–122.

Hamed Taherdoost. 2016. Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *International journal of academic research in management (IJARM)*, 5.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. 2025. JudgeBench: A Benchmark for Evaluating LLM-Based Judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Zichen Tang, Jiacheng Liu, Zhongjun Yang, Rongjin Li, Zihua Rong, Haoyang He, Zhuodi Hao, Xinyang Hu, Kun Ji, Ziyan Ma, and 1 others. 2025. FinMMR: Make Financial Numerical Reasoning More Multimodal, Comprehensive, and Challenging. *arXiv preprint arXiv:2508.04625*.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown. Accessed:2024-11-28.

Simone Teufel and Marc Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational linguistics*, 28(4):409–445.

Vasilis Theoharakis, Chris Voss, George C Hadjinicola, and Andreas C Soteriou. 2007. Insights into factors affecting production and operations management (pom) journal evaluation. *Journal of Operations Management*, 25(4):932–955.

Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Yunjie Ji, Han Zhao, and Xiangang Li. 2025. DeepDistill: Enhancing LLM Reasoning Capabilities via Large-Scale Difficulty-Graded Data Training. *arXiv preprint arXiv:2504.17565*.

Sarah Vahed. 2025. Interdisciplinarity opens new frontiers for decision science: Interdisciplinary research. *Nature Reviews Psychology*, 4(8):505–505.

Steven Vajda. 2009. *Mathematical programming*. Courier Corporation.

Yuxuan Wan, Wenxuan Wang, Yiliu Yang, Youliang Yuan, Jen-tse Huang, Pinjia He, Wenxiang Jiao, and Michael Lyu. 2024. LogicAsker: Evaluating and Improving the Logical Reasoning Ability of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2124–2155.

Bin Wang, Zhuangcheng Gu, Guang Liang, Chao Xu, Bo Zhang, Botian Shi, and Conghui He. 2024a. Unimernet: A Universal Network for Real-world Mathematical Expression Recognition. *arXiv preprint arXiv:2404.15254*.

Bin Wang, Fan Wu, Linke Ouyang, Zhuangcheng Gu, Rui Zhang, Renqiu Xia, Bo Zhang, and Conghui He. 2024b. CDM: A Reliable Metric for Fair and Accurate Formula Recognition Evaluation. *arXiv preprint arXiv:2409.03643*, 5(6).

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024c. MinerU: An Open-Source Solution for Precise Document Content Extraction. *arXiv preprint arXiv:2409.18839*.

Shuting Wang, Jiejun Tan, Zhicheng Dou, and Ji-Rong Wen. 2025. OmniEval: An Omnidirectional and Automatic RAG Evaluation Benchmark in Financial Domain. In *Proceedings of the 2025 conference on empirical methods in natural language processing*, pages 5737–5762.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in neural information processing systems*, volume 35, pages 24824–24837.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid Deep Sequential Modeling for Social Text-Driven Stock Prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1627–1630. ACM.

Yuhe Wu, Yuran Chen, Zhuang Liu, and Wayne Lin. 2025. Enhancing Financial Decision-making under Cyber Threats: a Dual-branch Framework Integrating Bayesian Deep Learning and Explainable AI. *Annals of Operations Research*, pages 1–33.

xAI. 2025. Models. Accessed:2025-07-09.

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. In *Advances in Neural Information Processing Systems*, volume 37, pages 95716–95743.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A Comprehensive Benchmark, Instruction Dataset and Large Language Model for Finance. In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484.

Jiang Xinguang, Hu Sihan, Yu Dingfu, Zhang Yuhao, Yang Zhongliang, Li Yu, Zhou Linna, and Valuesimplex AI Lab. 2023. FinLongEval.

Liang Xu, Lei Zhu, Yaotong Wu, and Hang Xue. 2024. SuperCLUE-Fin: Graded Fine-Grained Analysis of Chinese LLMs on Diverse Financial Tasks and Applications. *arXiv preprint arXiv:2404.19063*.

Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu, and Hongyuan Mei. 2025. FAMMA: A Benchmark for Financial Multilingual Multimodal Question Answering. *arXiv preprint arXiv:2410.04526*.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating Plausible Counterfactual Explanations for Deep Transformers in Financial Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6150–6160.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics.

Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. XFinBench: Benchmarking LLMs in Complex Financial Problem Solving and Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8715–8758.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024a. FinanceMATH: Knowledge-Intensive Math Reasoning in Finance Domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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 Long and Specialized Documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in neural information processing systems*, volume 36, pages 46595–46623.

Yaxian Zhou, Yufei Yuan, Kai Huang, and Xiangpei Hu. 2024. Can ChatGPT Perform a Grounded Theory Approach to Do Risk Analysis? An Empirical Study. *Journal of Management Information Systems*, 41(4):982–1015.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the Event: Corporate Events Detection for News-Based Event-Driven Trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124. Association for Computational Linguistics.

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 *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 3277–3287.

Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. 2024. Are Large Language Models Good Statisticians? In *Advances in Neural Information Processing Systems*, volume 37, pages 62697–62731.

A Benchmark Comparison

Table 5 compares BizCompass with existing benchmarks across both knowledge coverage and application dimensions. Most prior datasets focus narrowly on finance-related question answering or isolated analytical tasks, lacking broader domain representation and role-specific evaluation. In contrast, BizCompass achieves comprehensive disciplinary coverage spanning finance, economics, statistics, and operations management, while also incorporating three professional perspectives: analysis, trading, and consulting. Furthermore, it unifies multiple question formats, including single-choice, multiple-choice, and open-ended tasks, enabling a more systematic and fine-grained assessment of business reasoning capabilities in LLMs.

B Data Sources

Economics. Economics departments favour outlets that jointly signal methodological rigour, disciplinary reach, and balanced field coverage. Bibliometric surveys place the "Top Five" general journals—*AER*, *QJE*, *JPE*, *Econometrica*, and *RES*—at the centre of citation influence and promotion decisions (Kalaitzidakis et al., 2011; Heckman and Moktan, 2020). To avoid overweighting these flagships, we add the top-ranked field titles that remain in the overall top-30 after field normalisation—*JME*, *JIE*, *JDE*, *JHR*, and *JOPE*—as identified by multi-criteria and "ambition-adjusted" rankings (Combes and Linnemer, 2010; Engemann and Wall, 2009; Kodrzycki and Yu, 2006). Including *Experimental Economics* secures coverage of laboratory and behavioural methods (Cloos et al., 2023). Meta-science work shows that papers in these venues receive markedly more citations than the discipline average, ensuring durable visibility (Hamermesh, 2018). The selected core Economics Journal details are shown in Table 6.

Finance. Finance scholarship also values outlets that deliver methodological rigour, cross-field impact, and field balance. Citation and promotion studies show that the discipline's "Top Three" general journals—*Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*—dominate career-making citations (Bajo et al., 2020; Currie and Pandher, 2020; Borokhovich et al., 2000). To provide breadth, we add leading field journals that rank A/A$^+$ in active-scholar assessments (e.g. *Journal of Corporate Finance*, *Review of Finance*, *Review of Asset Pricing Studies*, *Journal of Financial Intermediation*, *Journal of Financial Stability*, *Journal of Financial Markets*, *Journal of International Money and Finance*) together with quantitative-methods, insurance/actuarial, and real-asset titles. Combined, the 30-journal set captures the citation core of contemporary financial economics (details are shown in Table 7).

Statistics. Citation capital in statistics is highly concentrated. Harter and Luby (Stigler, 1994) show that the ten most-cited statistics journals capture just over half ($\approx$ 52%) of all within-field citations, while Ruiz-Castillo and Waltman (Ruiz-Castillo and Waltman, 2015) report a journal-level Gini coefficient of ≈ 0.64—well above economics' 0.47. The American Statistical Association's tenure guidelines acknowledge this skew by naming only a handful of core venues (Association et al., 2024). Our eight-journal list—covering theory (*Biometrika*, *JRSS B*), computation/visualisation (*CS&DA*, *JCGS*), high-dimensional inference (*JMVA*), public-health methods (*Biostatistics*), policy applications (*JRSS C*), and a generalist forum (*Scand. J. Stat.*)—thus captures the field's reputational and citation nucleus without diluting prestige. The detailed Statistics journal sources are shown in Table 8.

Operations Management. OM/OR research is organized around a compact core. A Web of Science–based analysis for 2018–2022 by Sodhi and Tang shows that four INFORMS titles, namely *Management Science*, *Operations Research*, *Transportation Science*, and *INFORMS Journal on Applied Analytics* (formerly *Interfaces*), collectively account for $\approx$ 51% of all citations in the domain (Theoharakis et al., 2007). These journals hold high-tier AJG/ABS-2021 ratings and feature prominently in Harzing's Journal Quality List (Hudson, 2025; Edition, 2000).

Our initial source set therefore comprises these INFORMS anchors together with widely recognized OM outlets from Wiley (*Decision Sciences*; *Journal of Operations Management*; *Naval Research Logistics*), Taylor & Francis (*IISE Transactions*; *International Journal of Production Research*), Springer (*Journal of Optimization Theory and Applications*; *Journal of Scheduling*; *Mathematical Programming*; *Mathematical Programming Computation*), and Elsevier (*Operations Research Letters*). Inclusion is justified by standing in the AJG/ABS-2021 guide (typically $\geq$3 in Opera-

Benchmark	Knowledge Coverage				Application Dimensions			Question Types		
	Finance	Economics	Statistics	OM	Analysis	Trading	Consulting	SC.	MC.	OQ.
FinQA (Chen et al., 2021a)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
TAT-QA (Zhu et al., 2021)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
CovFinQA (Chen et al., 2022)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
FLUE (Shah et al., 2022)	✓	✓	✗	✗	✓	✗	✗	✓	✗	✗
FinanceBench (Islam et al., 2023)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
FinEval (Guo et al., 2025)	✓	✓	✗	✗	✓	✗	✓	✓	✗	✓
BBT-CFLEB (Lu et al., 2023)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
FinLBench (Xinguang et al., 2023)	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓
PIXIU (Xie et al., 2023)	✓	✓	✗	✗	✓	✓	✗	✗	✗	✓
DISC-FinLLM (Chen et al., 2023)	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
BBT-Fin (Lu et al., 2023)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
DOCMATH-EVAL (Zhao et al., 2024b)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
StatQA (Zhu et al., 2024)	✗	✗	✓	✗	✓	✗	✗	✗	✗	✓
QRDATA (Liu et al., 2024)	✗	✗	✓	✗	✓	✗	✗	✗	✓	✓
BizBench (Krumdick et al., 2024)	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓
FinBen (Xie et al., 2024)	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓
FinanceMATH (Zhao et al., 2024a)	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗
SuperCluE-Fin (Xu et al., 2024)	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
FinDABench (Liu et al., 2025)	✓	✗	✓	✗	✓	✗	✗	✗	✗	✓
AlphaFin (Li et al., 2024b)	✓	✗	✗	✗	✓	✓	✓	✗	✗	✓
FinTextQA (Chen et al., 2024)	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓
OmniEval (Wang et al., 2025)	✓	✗	✗	✓	✓	✗	✓	✗	✗	✓
INS-MMBench (Lin et al., 2025)	✗	✗	✗	✓	✓	✗	✗	✗	✓	✓
FinanceQA (Mateega et al., 2025)	✓	✗	✗	✗	✓	✗	✗	✗	✗	✓
XFinBench (Zhang et al., 2025)	✓	✓	✗	✗	✓	✗	✗	✓	✗	✓
Finance Agent (Bigeard et al., 2025)	✓	✗	✗	✗	✓	✗	✗	✓	✗	✓
CFinBench (Nie et al., 2025)	✓	✓	✗	✗	✓	✗	✗	✓	✓	✗
FinMaster (Jiang et al., 2025)	✓	✓	✓	✓	✓	✗	✓	✗	✗	✓
FinMMR (Tang et al., 2025)	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗
StatLLM (Song et al., 2026)	✓	✗	✓	✗	✓	✗	✗	✗	✗	✓
SECQUE (BenYoash et al., 2025)	✓	✗	✓	✓	✓	✗	✗	✗	✗	✓
FAMMA (Xue et al., 2025)	✓	✓	✗	✗	✓	✗	✗	✗	✓	✓
IndustryOR (Huang et al., 2025)	✗	✗	✗	✓	✓	✗	✓	✗	✗	✓
BizCompass (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 5: Comparison of existing benchmarks with BizCompass. Red ✓ indicates full coverage, and green ✗ indicates lack of coverage. "SC." is short for "Single Choice", "MC." is short for "Multiple Choice", and "OQ." is short for "Open Question". BizCompass achieves comprehensive disciplinary coverage and role-aware task design across analysis, trading, and consulting, while prior benchmarks are limited to narrower domains or tasks.

tions & Technology), consistent placement in JQL, and sponsorship by leading professional societies (e.g., the Mathematical Optimization Society for *Mathematical Programming*; the Institute of Industrial and Systems Engineers for *IISE Transactions*) (Vajda, 2009; Ding, 2023). The detailed Operation Management journal sources are shown in Table 9.

C Benchmark Construction

C.1 Phase1: Data Preprocessing

Document Conversion. We convert all raw material from PDF to Markdown using MinerU (Wang et al., 2024c; He et al., 2024). MinerU integrates layout-aware parsing (DocLayout-YOLO) and OCR (UniMERNet) (Wang et al., 2024a,b) to preserve structural elements such as section headings, mathematical formulas, tables, and inline expressions.

Data Cleaning. We pre-clean the converted Markdown raw material to obtain the analysis-ready corpus. First, we perform macro deletion: remove sections unrelated to question construction (Abstract, Related Work, Acknowledgements, Funding, Author/Ethics statements, References, and non-core appendices) and delete all figure content and containers. Second, we sanitize the remaining text: strip in-text citation markers, footnotes, and links while preserving the original wording, equations, tables, and section order; discard paragraphs with insufficient content or pedagogical

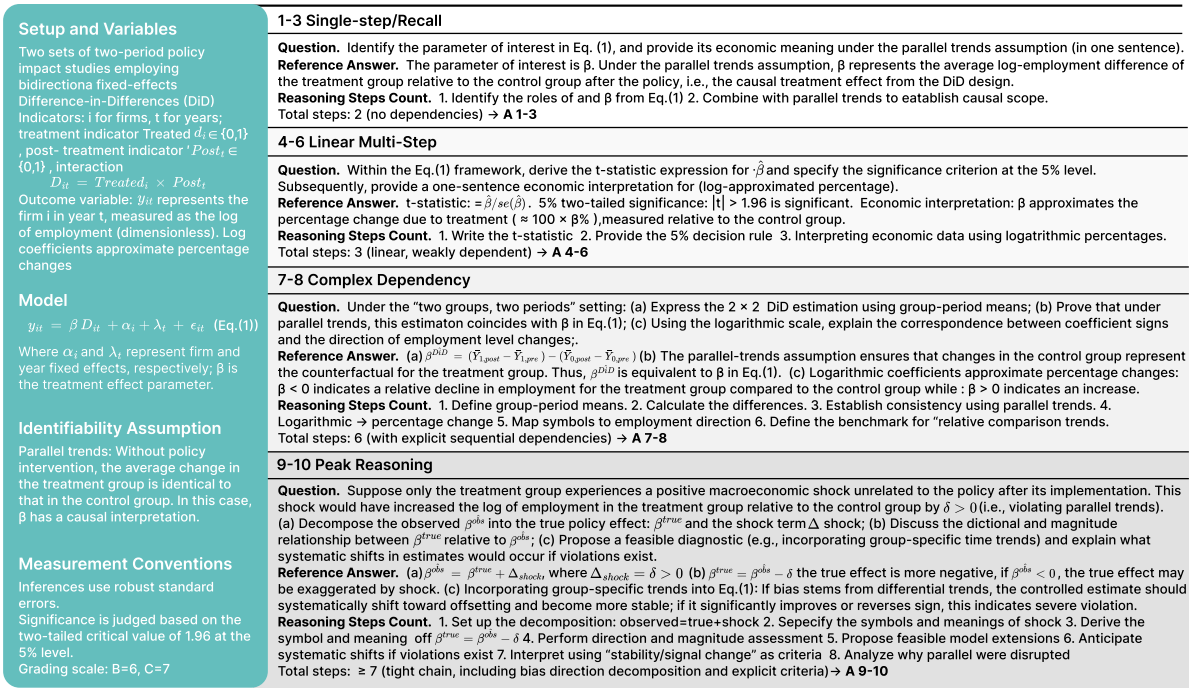

Figure 6: 4 examples showing different scores in A (Depth of Reasoning) dimension.

value.

C.2 Phase2: Question Construction

Due to the limited long-context understanding ability of LLMs, our experiments show that directly generating high-quality benchmark samples is impractical, as it exposes three major problems: the question stem and answering prerequisites are not closed, so key conditions are hard to specify in a single pass, largely because models are optimized to answer rather than to question incomplete premises and struggle to maintain reasoning consistency over long documents (Li et al., 2024a; Bai et al., 2024); answers are not traceable to explicit evidence, and cross-references can induce out-of-text inference; and the lack of auditable intermediate artifacts makes errors hard to localize and reproduce. Therefore, we adopt a six-step pipeline that decomposes construction into independent and verifiable stages. Each stage addresses one decision point, relies only on explicit evidence from the corpus, and outputs traceable intermediates and logs. This design preserves efficiency while improving item quality.

Section Summarization. To convert an unstructured corpus into structured evaluation units, this stage constructs knowledge units. First, we extract only three types of entities that are essential to a paper's core argument: key equations (e.g., identification equations and equilibrium conditions), core propositions (e.g., theorems and hypotheses), and key tables. The extraction rules are explicitly defined; for example, tables must report baseline regression results or key identification test outcomes. Subsequently, the process aggregates these entities according to a Holistic Integrity Principle. When a proposition and its associated key equation together form a self-contained logical unit, they are integrated into a macro knowledge point in order to preserve the complete argumentative structure. Finally, each atomic or macro knowledge point is assigned one or more labels from a predefined set of question types. By combining the pattern recognition capabilities of the LLM with explicit rules, we ensure that the extraction of knowledge units is both traceable and faithful to the original logical structure of the paper.

Question Decomposition. To assess integrative reasoning beyond factual recall, the core task at this stage is to synthesize isolated knowledge points into complex questions. We employ a process that uses the Question Type labels of a knowledge point as input to generate a synthetic instruction. This instruction directs the LLM to integrate the analytical dimensions implied by all associated labels and produce an initial question draft. For example, if a knowledge point is tagged with Methodology and Boundary Condition, the resulting instruction

Full title	Website
American Economic Association	
American Economic Journal: Applied Economics	*aeaweb.org/journals/app*
American Economic Journal: Economic Policy	*aeaweb.org/journals/pol*
American Economic Journal: Macroeconomics	*aeaweb.org/journals/mac*
American Economic Journal: Microeconomics	*aeaweb.org/journals/mic*
American Economic Review (incl. Papers & Proceedings)	*aeaweb.org/journals/aer*
Springer Nature	
Experimental Economics	*springer.com/journal/10683*
Journal of Economic Growth	*springer.com/journal/10887*
Oxford University Press	
The Economic Journal	*academic.oup.com/ej*
Journal of the European Economic Association	*academic.oup.com/jeea*
Quarterly Journal of Economics	*academic.oup.com/qje*
Review of Economic Studies	*academic.oup.com/restud*
Elsevier	
Journal of Development Economics	*sciencedirect.com/j/develop-econ*
Journal of Economic Theory	*sciencedirect.com/j/econ-theory*
Journal of International Economics	*sciencedirect.com/j/intl-econ*
Journal of Monetary Economics	*sciencedirect.com/j/monetary-econ*
Journal of Econometrics	*sciencedirect.com/j/econometrics*
Journal of Public Economics	*sciencedirect.com/j/public-econ*
Review of Economic Dynamics	*sciencedirect.com/j/rev-econ-dynam*
University of Wisconsin Press	
Journal of Human Resources	*jhr.uwpress.org*
University of Chicago Press	
Journal of Labor Economics	*journals.uchicago.edu/j/jole*
Journal of Political Economy	*journals.uchicago.edu/j/jpe*
Wiley–Blackwell	
Quantitative Economics	*qeconomics.org*
Econometrica	*econometricsociety.org/econometrica*
Theoretical Economics	*econtheory.org*
RAND Journal of Economics	*wiley.com/journal/17562171*
International Economic Review	*wiley.com/journal/14682354*
Journal of Applied Econometrics	*wiley.com/journal/10991255*
Cambridge University Press	
Econometric Theory	*cambridge.org/j/econometric-theory*
MIT Press	
Review of Economics and Statistics	*direct.mit.edu/rest*

Table 6: Core Economics Journals Selected as Sources for the BizCompass: Publishers and URLs

guides the LLM to formulate a question that requires explaining the method while also discussing its underlying assumptions. This approach shifts the assessment from factual recall to the ability to integrate knowledge. By anchoring question generation to traceable type labels, we enhance the transparency of intent in the construction of each question.

Retrieval & Backfilling. To ensure each question is self-contained, this stage's core task is to ground the query draft in the source text. We employ an in-context retrieval strategy, leveraging the long context window of Gemini 2.5 Pro. This LLM-as-Retriever approach is chosen over traditional chunk-based Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) to mitigate

Full title	Website
WILEY–BLACKWELL	
European Financial Management	*wiley.com/journal/1468036x*
Journal of Finance	*wiley.com/journal/15406261*
Journal of Risk and Insurance	*wiley.com/journal/15396975*
Mathematical Finance	*wiley.com/journal/14679965*
Real Estate Economics	*wiley.com/journal/15406229*
Financial Review	*wiley.com/journal/15406295*
International Review of Finance	*wiley.com/journal/14682443*
Journal of Financial Research	*wiley.com/journal/14756803*
Journal of Applied Corporate Finance	*wiley.com/journal/17456622*
Corporate Governance: An International Review	*wiley.com/journal/14678683*
ELSEVIER	
Journal of Corporate Finance	*elsevier.com/j/corp-finance*
Journal of Financial Intermediation	*elsevier.com/j/fin-intermed*
Journal of Financial Markets	*elsevier.com/j/fin-markets*
Journal of Financial Stability	*elsevier.com/j/fin-stability*
Insurance: Mathematics and Economics	*elsevier.com/j/ins-math-econ*
Journal of Financial Economics	*elsevier.com/j/fin-econ*
Journal of International Money and Finance	*elsevier.com/j/intl-money-fin*
SPRINGER NATURE	
Journal of Financial Services Research	*springer.com/journal/10693*
Journal of Real Estate Finance and Economics	*springer.com/journal/11146*
Review of Quantitative Finance and Accounting	*springer.com/journal/11156*
TAYLOR & FRANCIS	
Applied Financial Economics	*tandfonline.com/toc/rafe20*
European Journal of Finance	*tandfonline.com/toc/rejf20*
North American Actuarial Journal	*tandfonline.com/toc/uaaj20*
Scandinavian Actuarial Journal	*tandfonline.com/toc/sajf20*
CAMBRIDGE UNIVERSITY PRESS FOR THE INT. ACTUARIAL ASSOCIATION	
ASTIN Bulletin	*cambridge.org/j/astin-bulletin*
OXFORD UNIVERSITY PRESS	
Review of Financial Studies	*academic.oup.com/rfs*
Review of Finance	*academic.oup.com/rof*
Review of Asset Pricing Studies	*academic.oup.com/raps*
Review of Corporate Finance Studies	*academic.oup.com/rcfs*
Journal of Financial Econometrics	*academic.oup.com/jfec*

Table 7: Core Finance Journals Selected as Sources for the BizCompass: Publishers and URLs

known issues such as information fragmentation and the difficulty in capturing long-range dependencies (Sarthi et al., 2024; Han et al., 2024). By providing the full document as context, the model first locates and extracts verbatim the core passage that directly answers the question, which constitutes the answer draft. Second, it synthesizes a comprehensive set of evidence, including all necessary background like variable definitions, model assumptions, and key context. Adherence to a zero-paraphrasing principle is strictly enforced through prompting and automated checks. By constructing a ***Question-Answer-Evidence*** triplet, this stage ensures the verifiability of every question, establishing a foundation for reliable downstream evaluation.

Restructure. To transform discrete *Evidence* sets into structured units, our process begins by mapping each *Evidence* set to a standardized template: descriptive context and variable definitions are organized into the *Background*, core equations and data are placed in the *Model* section, and the initial

Full title	Website
OXFORD UNIVERSITY PRESS	
Biometrika	*academic.oup.com/biomet*
Biostatistics	*academic.oup.com/biostatistics*
JRSS Series B (Statistical Methodology)	*academic.oup.com/jrsssb*
JRSS Series C (Applied Statistics)	*academic.oup.com/jrssc*
ELSEVIER	
Computational Statistics & Data Analysis	*elsevier.com/j/comp-stat-data*
Journal of Multivariate Analysis	*elsevier.com/j/multivar-anlys*
TAYLOR & FRANCIS	
Journal of Computational and Graphical Statistics	*tandfonline.com/toc/ucgs20*
Journal of Business & Economic Statistics	*tandfonline.com/toc/ubes20*
WILEY–BLACKWELL	
Scandinavian Journal of Statistics	*wiley.com/journal/14679469*

Table 8: Core Statistics Journals Selected as Sources for the BizCompass: Publishers and URLs

Full title	Website
INFORMS	
Management Science	*pubsonline.informs.org/j/mnsc*
Operations Research	*pubsonline.informs.org/j/opre*
Transportation Science	*pubsonline.informs.org/j/trsc*
INFORMS Journal on Applied Analytics (formerly *Interfaces*)	*pubsonline.informs.org/j/inte*
WILEY	
Decision Sciences	*wiley.com/journal/15405915*
Journal of Operations Management	*wiley.com/journal/18731317*
Naval Research Logistics	*wiley.com/journal/15206750*
TAYLOR & FRANCIS	
IISE Transactions (formerly *IIE Transactions*)	*tandfonline.com/toc/uiie20*
International Journal of Production Research	*tandfonline.com/toc/tprs20*
SPRINGER	
Journal of Optimization Theory and Applications	*springer.com/journal/10957*
Journal of Scheduling	*springer.com/journal/10951*
Mathematical Programming	*springer.com/journal/10107*
Mathematical Programming Computation	*springer.com/journal/12532*
ELSEVIER	
Operations Research Letters	*elsevier.com/j/op-res-letters*

Table 9: Core Operations Management Journals Selected as Sources for the BizCompass: Publishers and URLs

question draft serves as the basis of the *Question*.

To enrich the processed cases with greater cognitive depth, we design executable textual constraints requiring that each question integrate multiple pieces of evidence and that its sub-questions follow a logical progression. To assess higher-order synthesis, the process further merges complementary cases into a more comprehensive question that forms a complete analytical chain, thereby simulating the multi-step reasoning required in actual research.

Scoring & Filtering. To ensure the quality of the question set, all generated questions must pass a data-driven filtering procedure that evaluates them using a quantitative scoring function across three dimensions: Reasoning Depth, Knowledge Synthesis, and Conceptual Centrality (see Appendix D). This filtering step helps maintain quality consistency

Source Category	AP		MTA			RM		SD		PM		AT	FD		FDA	
	A	T	A	T	C	A	C	T	C	A	C	T	A	C	A	C
Practitioner-oriented Textbooks	✓	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗
Business Documents	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗
Existing Benchmarks	✗	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✓	✓

Sources for ticks: AP–T: CIKM18 (Wu et al., 2018); MTA–A: FPB (Malo et al., 2014); MTA–T: M&A (Yang et al., 2020); MTA–C: FOMC (Shah et al., 2023); FD–A: CCFraud (Feng et al., 2023); FD–C: Taiwan (Feng et al., 2023); FDA–A: FNXL (Sharma et al., 2023), FinQA (Chen et al., 2021b), TATQA (Zhu et al., 2021), FinRED (Sharma et al., 2022); FDA–C: EDTSUM (Zhou et al., 2021).

Table 10: Application-Based Tasks across four source categories. AP=Asset Pricing; MTA=Market Trend Analysis; RM=Risk Management; SD=Strategy Development; PM=Portfolio Management; AT=Algorithmic Trading; FD=Fraud Detection; FDA=Financial Document Analysis. Subcolumns A/T/C denote Analysis/Trading/Consulting. (✓) indicates the task–role pair *uses data* from the given source category; (✗) indicates it *does not use* data from that source.

and reduces the workload of subsequent manual evaluation.

Type Conversion. z This stage optimizes the assessment format by matching the most suitable question type to each evaluation objective. A question is converted into a *Choice Question* when its answer space is relatively convergent and common fallacies can be used to construct plausible distractors; otherwise, it is retained as a *QA* problem. For choice questions, we leverage a fallacy knowledge base to construct distractors, as described in Appendix E. This process pairs the depth of QA-style reasoning with the efficiency and objectivity of choice questions (Haladyna, 2004; Ozuru et al., 2013).

C.3 Phase3: Evaluation

Framework Overview. We propose an expert-driven *Dual-Track Evaluation Strategy* for the two task families in the benchmark. The strategy focuses on item quality and proceeds in two stages: (i) Stratified expert sampling and recruitment; (ii) Dual-Track Multi-Dimensional Evaluation and Filtering.

Expert Recruitment. We use stratified purposive sampling to form two non-overlapping panels (Stratton, 2024), each confined to its track: (i) *Specialist Advisory Panel* for Knowledge-Based items ($n = 40$, equally split across Operations Management, Statistics, Finance, and Economics). Members hold at least a Master's degree, with 57.5% holding or currently pursuing a Ph.D. in their respective fields. (ii) *Core Practitioner Panel* for Application-Based items ($n = 30$, equally split across Analyst, Trader, and Consultant roles). Members possess a relevant degree and at least three years of verifiable full-time industry experience. Candidates were sourced from universities, professional communities, certification rosters (*e.g.*, CFA/FRM), and alumni networks of leading finance and consulting firms.

Evaluation and Filtering Rules. To implement our dual-track validation framework and ensure robustness, we adopted a dual-review coverage design, ensuring that every candidate item was independently adjudicated by at least two experts. To prevent cognitive fatigue, items were distributed in batches over a 2-week review period. Specifically, each expert in the knowledge-based track reviewed approximately 400 items, while each expert in the application-based track reviewed approximately 600 items. During the review, experts provided independent scores against predefined quality dimensions. To handle disagreements, we implemented a strict *"Review-Discuss-Resolve"* protocol. Items receiving conflicting decisions or low-quality scores were flagged as "discrepant items". Experts were required to discuss and decide on the final label. If consensus could not be reached after two rounds of discussion, the item was strictly discarded to maintain the high quality of the final dataset. The following sections detail the specific evaluation dimensions that underlie this review process:

Track A: Knowledge-Based Tasks. This track ensures that knowledge items measure LLMs' mastery of the theoretical foundations in Operations Management, Statistics, Finance, and Economics, with attention to core concepts.

- **Clarity.** Clarity is the basis of valid assessment. In psychometrics, ambiguity introduces construct-irrelevant variance, meaning scores are

influenced by factors unrelated to the target construct (e.g., general language decoding), which harms reliability and validity (Bandalos, 2018). In LLM benchmarking, an unclear item probes guessing under vague instructions rather than domain knowledge. Ensuring clarity is the first step toward face validity, fixing the semantic ground on which answers are given and allowing us to isolate the model's domain knowledge (Taherdoost, 2016; Wu et al., 2025). ***Core Question:*** *Is the problem statement precise and unambiguous?*

- **Conceptual centrality.** Research in cognitive science shows that domain knowledge forms an interdependent network rather than a set of isolated facts (Sloman, 1993). A concept's centrality depends on how many other concepts rely on it. Scientific discourse analyses also find that sentences tied to central arguments are more informative and salient (Teufel and Moens, 2002). Items that focus on core concepts, such as opportunity cost in economics or the central limit theorem in statistics, carrying greater diagnostic value than those focused on peripheral facts. They more effectively assess whether a model has internalized the principles that drive reasoning, rather than merely memorized fragments. ***Core Question:*** *Does the item test a core concept of the discipline?*

- **Domain representativeness.** This dimension instantiates content validity: the extent to which an instrument covers the important parts of the domain (Haynes et al., 1995; Taherdoost, 2016). A benchmark with strong content validity samples items that represent the field's overall structure. This reduces "gaming the benchmark," where a model overfits narrow subtopics without global understanding. Expert judgments on representativeness ensure breadth and balance so the benchmark serves as a valid proxy for composite competence. ***Core Question:*** *Does the item reflect knowledge typically expected of practitioners in this field?*

- **Cognitive complexity.** Cognitive complexity evaluates the depth and level of processes needed to solve the item. Following Bloom's Revised Taxonomy, processes span remember, understand, apply, analyze, evaluate, and create (Anderson and Krathwohl, 2001). A high-quality benchmark includes a sufficient share of items that require higher-order processes (e.g., analysis, evaluation) to differentiate reasoning ability (Rush et al., 2016). This dimension filters recall-only items and ensures the bank elicits and measures multi-step reasoning, concept integration, and critical judgment rather than mere information retrieval. ***Core Question:*** *What level of cognition is required, from recall to multi-step reasoning?*

Track B: Application-Based Tasks. This track ensures that application items reflect realistic business conditions and align with the core competencies of specific roles.

- **Clarity.** As in Track A, clarity is essential. An unclear business scenario forces the model to assume missing conditions, turning the task from a reasoning test into a guessing exercise. This undermines reliability—the reproducibility of results. Clear items allow us to attribute performance differences to underlying ability rather than arbitrary interpretations (Taherdoost, 2016). ***Core Question:*** *Are the scenario and task clearly described with explicit intent?*

- **Realism.** Realism evaluates consistency with actual business practice. It secures ecological validity, i.e., the degree to which results generalize to real, non-research decision contexts (Brunswik, 1955). An application item that ignores business logic, data constraints, or decision pressure cannot demonstrate practical utility even if a model produces a correct answer. Realism links benchmark performance to real-world potential. ***Core Question:*** *Does the item reflect a credible real-world business scenario?*

- **Role-relevance.** This dimension assesses alignment with role responsibilities using standardized Job/Task Analysis (JTA) to establish content validity (Lane et al., 2016). JTA is the gold standard in professional testing and certification: subject-matter experts identify and rate the tasks and knowledge essential to each role. Having role-matched experts judge relevance provides evidence for construct validity that the benchmark measures the claimed role-specific abilities. ***Core Question:*** *Is the item closely aligned with the core duties of the specified role?*

- **Difficulty.** Difficulty captures the depth of domain knowledge, the length and complexity of the reasoning chain, and the breadth of integration. Assessment theory requires sufficient dis-

criminatory power to separate ability levels (Lane et al., 2016). Expert difficulty judgments help build a well-graded bank, reducing ceiling and floor effects and enabling more precise ranking and evaluation across models. ***Core Question:*** *What level of combined knowledge, reasoning, and information integration is required?*

D Question Quality Filter

D.1 Sample Quality Evaluation Dimensions

We argue that a good benchmark sample should exhibit sufficient reasoning depth, integrate information comprehensively, and remain closely aligned with the primary claim of raw data. Correspondingly, we propose three dimensions to systematically evaluate sample quality, namely A (Depth of Reasoning), B (Degree of Synthesis), and C (Conceptual Centrality), which we define as follows.

Depth of Reasoning. Depth of Reasoning captures whether an example goes beyond surface recall to require multi-step inference, making it crucial for distinguishing genuine reasoning ability in LLMs. Classic benchmarks such as GSM8K probe this capacity through stepwise mathematical reasoning (Cobbe et al., 2021), and subsequent work on chain-of-thought prompting has shown that explicit intermediate reasoning steps are essential for solving complex problems, underscoring the importance of evaluating reasoning depth (Wei et al., 2022b). The criteria for different scoring bands are classified in Table 11. To elucidate the distinctions among the four different scoring bands, Figure 6 illustrates representative questions and corresponding reasoning examples derived from the same background information.

Degree of Synthesis. Degree of Synthesis captures the extent to which an example requires integrating multiple pieces of information into a coherent whole. This dimension is essential because real-world problems often demand the fusion of knowledge across domains or sources, making synthesis a key determinant of benchmark quality. For instance, HotpotQA requires reasoning across multiple documents to synthesize evidence (Yang et al., 2018), while MultiRC involves combining information scattered across sentences (Khashabi et al., 2018).

Conceptual Centrality. We define conceptual centrality as the closeness of a text unit to the primary claim or contribution of a paper. Prior work in scientific discourse analysis has shown that sentences aligned with central arguments are typically more information-dense and salient than peripheral content (Teufel and Moens, 2002), and thus are better suited for generating higher-quality samples.

Operational Definitions. *Information source.* A source used directly by the question and the answer that determines the conclusion. Count a source only if removing it would change the answer. One or more of the following six types:

- *Equations or formal conditions* (model relations, constraints, optimality conditions)
- *Numerical tables* (coefficients, standard errors, sample sizes, R^2)
- *Identification or validity assumptions* (explicitly stated conditions)
- *Variable and parameter definitions* (meaning, units, transformations)
- *Context and sample design* (treatment definitions, timing, scope)
- *Estimation and design details* (procedures, settings, data construction)

Reasoning step. An inference step is a necessary operation that yields a new intermediate or final conclusion from the premises. Count one step if and only if at least one of the following conditions holds. Splitting is allowed only when an intermediate statement is necessary and substantive; do not count mere transcription, reading a single table entry, or mechanically breaking one algebraic manipulation into micro-steps:

- Deriving a new property or directional claim from a formal condition
- Combining two different information sources to obtain a joint conclusion
- Performing one statistical decision or interval judgment
- Conducting one comparative static, counterfactual, or required simplification within the model

D.2 Sample Quality Scoring

To ensure the quality of benchmark samples, we adopt a two-step filtering framework. It consists of *Multi-dimensional Evaluation*, aggregating the

three quality dimensions into a weighted score to capture overall adequacy, and *Dimension-Wise Filtering*, enforcing minimum thresholds on each dimension to prevent samples with severe deficiencies from being retained. We next describe how these two scoring procedures are aligned with expert annotations.

Operational Scoring Rules and Bands. Using the operational definitions of information source and reasoning step, we assign per-example scores on three dimensions *Depth of Reasoning* (A), *Degree of Synthesis* (B) and *Conceptual Centrality* (C). For each dimension, choose an integer from 1–10 by selecting the highest band fully satisfied by the example and provide a one-sentence justification. The band criteria are summarized in Table 11. The overall score is Eq. 1; weight estimation follows.

Multi-dimensional Evaluation. Absolute human scores are often subjective and inconsistent across annotators, making them unreliable as direct supervision. To address this, we adopt a preference-based approach: experts were asked to compare triplets of examples and provide pairwise preferences (e.g., $\text{example}_1 \succ \text{example}_2$). This yields a set of pairwise constraints $\mathcal{P} = \{(i, j)\}$, where (i, j) indicates that example i is preferred over j.

Given the weighted scoring function

$$S(x) = \alpha A(x) + \beta B(x) + \gamma C(x), \qquad (1)$$

we require that $S(x_i) > S(x_j)$ whenever $(i, j) \in \mathcal{P}$. To enforce this, we adopt a pairwise ranking loss from the learning-to-rank literature (Burges et al., 2005):

$$\mathcal{L}(\alpha, \beta, \gamma) = \sum_{(i,j)\in\mathcal{P}} \log\left(1 + \exp(-(S(x_i) - S(x_j)))\right).$$

Minimizing this loss is equivalent to maximizing the likelihood under a Bradley–Terry preference model, which assumes that the probability of preferring i over j increases with $S(x_i) - S(x_j)$.

In practice, we normalize the coefficients (α, β, γ) to sum to one, and optimize them using gradient-based methods over preference annotations. Specifically, 10 human experts provided pairwise preference labels on 500 triplets of benchmark examples, yielding a set of training constraints for the ranking model. This process yields the final weights $(0.4, 0.4, 0.2)$, indicating that *Depth of Reasoning* and *Degree of Synthesis* contribute most strongly to perceived sample quality, while *Conceptual Centrality* plays a supportive but non-negligible role.

Dimension-Wise Filtering. While the weighted score reflects relative sample quality, it does not directly guarantee the rejection of particularly poor cases. We therefore impose a filtering threshold. Let $\mathcal{R}$ denote the set of samples rejected by human experts, and let $\hat{\mathcal{R}}(T)$ denote the set of samples filtered out by the automatic rule with threshold T. The rejection recall is defined as:

$$\text{Recall}_{\text{reject}}(T) = \frac{|\mathcal{R} \cap \hat{\mathcal{R}}(T)|}{|\mathcal{R}|}.$$

This measures the proportion of expert-rejected samples that are also eliminated by the automatic threshold. When $T = 6.0$, the recall reaches 90%, showing that our criterion effectively recovers the majority of samples deemed unacceptable by human experts. In practice, we additionally require that each dimension score is at least 6.0, preventing examples with severe deficiencies in any single aspect from being retained.

E Assessment Type Conversion

E.1 Aligning Assessment Formats with Cognitive Constructs

Rationale for the Dual-Format Assessment Strategy. The choice of an assessment format is a fundamental issue of construct validity, centered on the alignment between the evaluation tool and the cognitive processes being measured, rather than mere scoring efficiency (Messick, 1995). Psychometric literature clearly distinguishes between two primary formats: *Constructed-Response (CR)* and *Selected-Response (SR)* (Lane et al., 2016). As a typical CR format, the open-ended Question-Answer (QA) assesses active, generative cognitive processes. In contrast, the Multiple-Choice Question (MCQ), a typical SR format, evaluates recognition-based processes that rely more heavily on prior knowledge and familiarity (Ozuru et al., 2013).

The core objective of BizCompass is to evaluate the advanced cognitive capabilities of LLMs within professional domains. Expertise in these domains, whether in academic research or business analysis, requires both generative abilities (e.g., composing in-depth analyses) and discriminative abilities (e.g., making precise judgments among viable options) (Shanteau, 1992). Consequently, a bench-

A. Depth of Reasoning	
1–3	*Steps:* $\leq$2; *Dependence:* none; *Ops:* no comparative statics or counterfactuals.
4–6	*Steps:* 3–4; *Dependence:* weak; *Ops:* at least one statistical decision or necessary calculation.
7–8	*Steps:* 5–6; *Dependence:* later conclusions rely on earlier results; *Ops:* at least one of comparative statics, counterfactual, or multi-stage derivation; *Sources:* connects at least two information sources.
9–10	*Steps:* $\geq$7 tightly linked; *Dependence:* strong chain; *Ops:* substantive setup change or bias direction/decomposition under violated assumptions with explicit criteria.
B. Degree of Synthesis	
1–3	Single source only.
4–6	Two distinct sources, both necessary and used substantively.
7–8	At least three sources used substantively, including an equation/formal condition and a numerical table.
9–10	More than three sources used substantively with non-obvious correspondence or mutual constraints; each source is necessary for the main conclusion.
C. Conceptual Centrality	
1–3	Peripheral to the paper's main claims.
4–6	Supportive but not core.
7–8	Directly serves a key identification condition, core mechanism, or baseline interpretation.
9–10	Core contribution (e.g., main theorem, baseline headline result, decisive mechanism evidence).

Table 11: Operational scoring bands for A–C dimensions.

mark relying on a single format would lack ecological validity, failing to comprehensively measure a model's integrated performance on the diverse tasks required in authentic, complex professional settings (Brunswik, 1955).

This distinction forms the theoretical cornerstone of our protocol. We treat the decision to convert a QA item to an MCQ as a *validity claim*: we posit that, for the specific assessment target, the core cognitive construct is more appropriately measured through discrimination among options rather than through unprompted generation (Messick, 1995). By strategically deploying both formats, our framework simulates a broader spectrum of real-world cognitive demands, thereby providing a more comprehensive and reliable assessment of an LLM's potential in professional domains.

Type Conversion Evaluation Dimensions. We argue that a rigorous format conversion decision requires a systematic evaluation of two core issues: the ability to losslessly capture the question's core construct in a multiple-choice format, and the potential for constructing distractors with high value. We therefore propose two dimensions to guide this evaluation.

- **Conceptual Clarity.** Conceptual Clarity is defined as the degree to which a question's core answer is convergent, determining if it can be losslessly captured by a single option. The clarity of the answer space is a cornerstone of assessment objectivity. If the answer space is divergent and permits multiple valid interpretations, the evaluation's focus shifts to the reasoning process rather than the final conclusion, which contradicts the fundamental purpose of multiple-choice questions to measure precise judgment (Rush et al., 2016).

- **Misconception Potential.** Misconception Potential captures whether a set of common fallacies exists for constructing distractors. The core value of this potential lies in enhancing a choice question's discriminability: its ability to effectively differentiate between examinees of varying proficiency levels. Commonsense reasoning benchmarks such as COSMOS QA demonstrate that high-quality distractors based on common errors are critical for building challenging multiple-choice questions (Huang et al., 2019). If a question does not permit the systematic construction of functional distractors, its multiple-choice version tends to degenerate into a simple pattern-matching test, thereby losing its diagnostic value (Taherdoost, 2016).

Type Conversion Scoring. To aggregate the scores from the two dimensions above into a single conversion suitability score, we employ the same preference-based learning methodology of Appendix D.2. This ensures a consistent, data-driven approach to weight determination. The re-

sulting decision function is:

$$\text{Total Score} = 0.5 \times A + 0.5 \times B.$$

where A and B represent the scores for Conceptual Clarity and Misconception Potential, respectively. An item is converted to the multiple-choice format only if its total score is at least 9.0; otherwise, it is retained in its original QA format. This high threshold ensures that only items demonstrating exceptional suitability on both dimensions are converted, thereby prioritizing the retention of QA problems that assess deep reasoning.

E.2 Framework for Distractor Design

The quality of distractors is an important factor of a multiple-choice question's measurement precision and diagnostic power. To mitigate the validity threats posed by arbitrarily created distractors, we adopt a evidence-based framework for distractor design. This framework is grounded in a core tenet of modern test theory that validity is not something checked after a test is complete, but rather something built into the instrument through a principled development process (Haladyna, 2004). It categorizes common errors into five core cognitive domains: conceptual, procedural, calculation, logic, and cognitive bias, as detailed in Table 12.

Formatting Rules. In addition to the content guidelines provided by the framework, all distractors must adhere to strict formatting rules to ensure fairness and prevent cueing. These include maintaining grammatical parallelism, ensuring options are mutually exclusive for single-choice questions, and avoiding the use of "all of the above" or "none of the above".

F Hyperparameter Selection

We conduct hyperparameter optimization for the two different question types to ensure optimal model performance. Specifically, we perform grid search on two key sampling parameters: temperature (ranging from 0 to 0.8) and top-p (ranging from 0.6 to 0.95).

For single-choice questions, we evaluate performance using accuracy as the metric, while for general QA tasks, we employ LLM evaluation scores. The grid search results are presented in Table 13, where the optimal hyperparameter combinations for each question type are highlighted. Based on these results, we select temperature=0.8 and top-p=0.95 for single-choice questions, and temperature=0.8 and top-p=0.6 for general QA tasks in our experiments.

G Practical Performance Metrics for Domain-Specific SFT

To assess the practical feasibility of deploying fine-tuned LLMs in business workflows, we profiled the computational cost of the fine-tuned model across different problem formats in BizCompass. The evaluated model was deployed on a server equipped with four NVIDIA A40 GPUs. We tracked four key performance indicators: Average Runtime, Time to First Token (TTFT), Token Generation Speed (Tokens/s), and Average Memory Usage (both GPU VRAM and system RAM). As summarized in Table 14, tasks requiring extensive context processing and generative reasoning, such as Table QA and General QA, demand significantly higher runtimes compared to discriminative tasks (Single and Multiple Choice). However, the underlying memory footprint remains largely stable across all question types, constrained primarily by the model's static weight allocation.

H Evaluation Results

Error Type	Definition and Cognitive Target
Conceptual Errors	
Common Misconception	Based on intuitive beliefs common among learners but contrary to scientific facts (Gierl et al., 2017; Radatz, 1979). *Target: Probe stable alternative concepts.*
Almost Correct	Mostly correct but missing one key detail (Haladyna et al., 2002). *Target: Separate precise understanding from surface recall.*
Opposite Idea	States the opposite of the correct idea (Gierl et al., 2017). *Target: Detect inverted understanding of a core idea.*
Procedural Errors	
Missing Step	Skips a required step; the distractor is the result after skipping (Radatz, 1979). *Target: Expose gaps in procedure or working memory.*
Wrong Order	All steps are present but in the wrong order (Radatz, 1979). *Target: Check control of procedure sequence.*
Rule Misuse	Uses the right procedure but misapplies a sub-rule or formula (Radatz, 1979). *Target: Locate weak links inside the procedure.*
Calculation Errors	
Sign Error	Flips the sign during calculation (Radatz, 1979). *Target: Check attention to sign rules.*
Unit Error	Uses the wrong unit or converts units incorrectly (Radatz, 1979). *Target: Check understanding of units and conversion.*
Order Error	Ignores the standard order of operations (Radatz, 1979). *Target: Find gaps in basic rules.*
Logical Errors	
Treating Correlation as Cause	Takes correlation as causation (Wan et al., 2024). *Target: Distinguish association from cause.*
Hasty Generalization	Concludes from too little or biased evidence (Wan et al., 2024). *Target: Check grasp of representativeness and inference.*
Invalid Logic	Uses an invalid form (e.g., affirming the consequent) (Wan et al., 2024). *Target: Evaluate logical soundness.*
Cognitive Bias Errors	
True but Irrelevant	Factually correct but irrelevant to the question (Forster and Lavie, 2008). *Target: Test filtering of relevant information.*
Anchor Trap	A number looks plausible due to an unrelated anchor in the stem (Berthet, 2022). *Target: Check influence of anchoring.*
Confirmation Trap	Fits prior belief but conflicts with given evidence (Nickerson, 1998; Berthet, 2022). *Target: Test ability to update belief.*

Table 12: Error Types Underlying Multiple-Choice Distractors

	Single Choice (Metric: Accuracy)					General QA (Metric: LLM Eval)				
	Temp.=0	Temp.=0.2	Temp.=0.4	Temp.=0.6	Temp.=0.8	Temp.=0	Temp.=0.2	Temp.=0.4	Temp.=0.6	Temp.=0.8
Top-p=0.6	32.73%	32.67%	32.67%	33.16%	34.92%	3.16	2.60	2.53	2.48	3.06
Top-p=0.8		33.16%	33.21%	34.92%	34.65%		2.49	2.56	2.52	2.98
Top-p=0.9		33.80%	33.80%	34.17%	34.01%		2.53	2.53	2.55	3.02
Top-p=0.95		32.67%	34.54%	33.96%	35.93%		2.59	2.57	2.49	3.01

Table 13: Hyperparameter grid search results. The selected hyperparameter combinations for Choice and QA question types have been highlighted.

Question Type	Avg Runtime (s)	Avg TTFT (s)	Tokens/s	Avg GPU Memory (MB)	Avg RAM (MB)
Table QA	80.238	0.018	6.014	15325.648	2434.830
General QA	83.441	0.011	6.031	15325.646	2542.343
Single Choice	0.872	0.008	5.238	15325.644	2615.969
Multiple Choice	1.079	0.008	4.233	15325.645	2575.021

Table 14: Computational cost and practical performance metrics of the fine-tuned model deployed on $4\times$ NVIDIA A40 GPUs across different question types.

Model	Size	Thinking	FIN				ECON			
			SC.	MC.	TableQA	GeneralQA	SC.	MC.	TableQA	GeneralQA
GPT-4o-2024-05-13	N/D	✗	69.14%	58.18% / 86.57%	4.841	4.736	58.79%	40.44% / 79.45%	4.741	4.665
GPT-4.1-2025-04-14	N/D	✗	76.55%	63.94% / 89.39%	4.950	4.957	67.88%	52.04% / 85.32%	4.950	4.955
GPT-5-2025-08-07	N/D	✓	89.20%	76.97% / 93.68%	4.982	4.986	88.79%	71.79% / 91.65%	4.967	4.992
Gemini-2.5-Flash-thinking	N/D	✓	89.82%	74.24% / 92.95%	4.942	4.903	83.33%	65.20% / 89.81%	4.907	4.919
Gemini-2.5-Pro-thinking	N/D	✓	91.05%	84.55% / 95.74%	4.954	4.948	88.79%	75.24% / 93.30%	4.953	4.940
Llama3-8B	8B	✗	38.89%	14.55% / 72.21%	3.240	2.770	30.61%	14.11% / 69.86%	2.961	2.452
Llama-3-70B-Instruct	70B	✗	65.13%	53.33% / 83.41%	4.603	4.432	56.36%	37.93% / 77.73%	4.443	4.272
Llama-3.3-70B-Instruct	70B	✗	71.61%	54.24% / 85.00%	4.799	4.688	62.42%	42.63% / 80.06%	4.693	4.616
Llama-4-Scout	17B	✗	76.24%	56.97% / 85.81%	4.128	3.656	60.00%	41.69% / 81.46%	3.905	3.295
Claude 3.7 Sonnet-20250219	N/D	✗	74.08%	62.73% / 87.88%	4.942	4.906	66.36%	48.59% / 82.86%	4.919	4.885
Claude 4 Sonnet-20250514	N/D	✓	88.27%	74.85% / 93.23%	4.939	4.896	82.12%	66.77% / 90.52%	4.895	4.870
Claude 4 Opus-20250514	N/D	✓	90.43%	81.21% / 94.98%	4.937	4.904	85.76%	77.74% / 93.72%	4.904	4.886
Claude 3.7 Sonnet-20250219-thinking	N/D	✓	90.13%	79.09% / 94.72%	4.942	4.892	85.15%	69.91% / 91.04%	4.925	4.882
Qwen3-235B-A22B	235B	✓	87.66%	75.76% / 93.80%	4.876	4.868	84.24%	72.73% / 92.05%	4.859	4.854
Qwen2.5-72B-Instruct	72B	✗	73.15%	53.64% / 87.89%	4.757	4.625	59.09%	42.32% / 81.03%	4.673	4.529
Qwen/QwQ-32B	32B	✓	88.27%	74.85% / 93.63%	4.870	4.801	82.12%	65.52% / 90.26%	4.848	4.688
Grok4-0709	N/D	✗	86.11%	80.00% / 94.85%	4.950	4.915	82.42%	70.53% / 92.22%	4.935	4.941
Grok3	N/D	✗	74.08%	65.45% / 89.35%	4.955	4.902	64.85%	50.47% / 85.32%	4.909	4.884
DeepSeek-V3-250324	671B	✗	79.02%	64.24% / 90.58%	4.891	4.828	68.49%	52.35% / 85.93%	4.846	4.776
DeepSeek-R1-0528	671B	✓	87.04%	76.36% / 93.53%	4.930	4.893	81.82%	65.52% / 90.32%	4.901	4.914
DeepSeek-R1-Distill-Qwen-7B	7B	✓	72.23%	50.61% / 84.61%	3.973	3.942	58.18%	35.42% / 77.04%	3.611	3.639
Deepseek-R1-Distill-Llama-70B	70B	✓	86.73%	73.33% / 92.63%	4.677	4.684	78.79%	61.13% / 87.16%	4.584	4.524
DeepSeek-R1-Distill-Qwen-32B	32B	✓	82.41%	71.21% / 92.44%	4.622	4.584	78.79%	56.74% / 86.50%	4.596	4.510

Model	Size	Thinking	OM				STAT			
			SC.	MC.	TableQA	GeneralQA	SC.	MC.	TableQA	GeneralQA
GPT-4o-2024-05-13	N/D	✗	56.35%	48.30% / 83.49%	4.665	4.603	58.49%	50.62% / 83.55%	4.816	4.693
GPT-4.1-2025-04-14	N/D	✗	66.88%	56.97% / 86.92%	4.945	4.962	69.50%	62.04% / 89.55%	4.972	4.980
GPT-5-2025-08-07	N/D	✓	86.38%	72.14% / 92.25%	4.972	4.986	90.88%	76.85% / 93.34%	4.985	4.992
Gemini-2.5-Flash-thinking	N/D	✓	83.90%	70.28% / 91.21%	4.913	4.851	89.62%	71.30% / 92.16%	4.955	4.918
Gemini-2.5-Pro-thinking	N/D	✓	87.62%	77.71% / 93.45%	4.940	4.932	92.14%	79.32% / 94.84%	4.955	4.975
Llama3-8B	8B	✗	31.27%	21.05% / 73.25%	2.916	2.729	29.87%	16.36% / 69.77%	3.351	2.693
Llama-3-70B-Instruct	70B	✗	56.66%	50.77% / 83.47%	4.384	4.253	59.43%	45.37% / 83.42%	4.564	4.384
Llama-3.3-70B-Instruct	70B	✗	64.71%	56.35% / 85.93%	4.657	4.591	66.98%	48.77% / 85.00%	4.775	4.665
Llama-4-Scout	17B	✗	58.52%	45.20% / 83.18%	3.486	3.230	65.09%	49.07% / 85.29%	3.850	3.385
Claude 3.7 Sonnet-20250219	N/D	✗	60.68%	56.04% / 85.29%	4.882	4.873	64.15%	54.01% / 86.24%	4.954	4.912
Claude 4 Sonnet-20250514	N/D	✓	82.35%	75.23% / 93.81%	4.894	4.860	89.62%	71.30% / 92.47%	4.938	4.873
Claude 4 Opus-20250514	N/D	✓	86.07%	74.30% / 93.10%	4.872	4.896	90.25%	79.01% / 93.87%	4.938	4.916
Claude 3.7 Sonnet-20250219-thinking	N/D	✓	84.52%	73.68% / 92.65%	4.920	4.861	90.57%	73.77% / 92.76%	4.936	4.899
Qwen3-235B-A22B	235B	✓	85.14%	74.92% / 93.42%	4.887	4.863	90.25%	74.07% / 93.68%	4.935	4.896
Qwen2.5-72B-Instruct	72B	✗	60.68%	52.63% / 86.66%	4.635	4.612	63.21%	49.69% / 86.88%	4.774	4.626
Qwen/QwQ-32B	32B	✓	83.90%	72.76% / 92.48%	4.758	4.648	89.94%	68.83% / 92.10%	4.900	4.771
Grok4-0709	N/D	✗	81.43%	71.83% / 92.33%	4.930	4.887	86.48%	70.68% / 92.83%	4.964	4.957
Grok3	N/D	✗	64.40%	56.35% / 87.57%	4.907	4.883	69.50%	60.19% / 88.79%	4.961	4.909
DeepSeek-V3-250324	671B	✗	67.18%	56.35% / 88.86%	4.835	4.807	76.73%	56.48% / 88.52%	4.863	4.856
DeepSeek-R1-0528	671B	✓	72.45%	69.66% / 92.00%	4.894	4.887	70.44%	71.30% / 91.63%	4.941	4.949
DeepSeek-R1-Distill-Qwen-7B	7B	✓	68.73%	44.27% / 83.27%	3.825	4.045	70.75%	40.43% / 83.05%	4.111	4.156
Deepseek-R1-Distill-Llama-70B	70B	✓	78.64%	65.63% / 89.32%	4.590	4.521	82.70%	61.42% / 88.45%	4.769	4.657
DeepSeek-R1-Distill-Qwen-32B	32B	✓	79.57%	65.02% / 90.76%	4.542	4.534	85.53%	61.11% / 89.42%	4.694	4.647

Table 15: Complete evaluation results for knowledge-based tasks in BizCompass. N/D means not disclosed.

Model	AP_A	CIKM18	FPB	M&A	FOMC	FRM	CCFraud	Taiwan	RM_C
	Accuracy								GPT Eval
GPT-4o-2024-05-13	57.98%	47.00%	97.20%	60.80%	63.31%	66.67%	67.40%	3.80%	4.272
GPT-4.1-2025-04-14	60.61%	46.60%	98.20%	57.80%	63.71%	67.81%	71.20%	3.60%	4.896
GPT-5-2025-08-07	96.16%	57.40%	99.40%	59.20%	66.33%	89.04%	85.00%	40.20%	4.940
Gemini-2.5-Flash-thinking	86.87%	51.40%	98.20%	67.20%	64.11%	70.55%	96.60%	37.20%	4.622
Gemini-2.5-Pro-thinking	84.85%	48.40%	95.00%	84.60%	65.32%	76.94%	80.20%	5.80%	4.798
Llama3-8B	24.65%	58.40%	87.00%	75.60%	59.27%	36.53%	6.60%	3.60%	2.998
Llama-3-70B-Instruct	65.25%	58.80%	97.40%	72.60%	68.35%	56.39%	7.60%	19.00%	3.902
Llama-3.3-70B-Instruct	72.73%	55.60%	93.20%	57.60%	69.96%	60.27%	12.80%	5.00%	4.182
Llama-4-Scout	61.21%	54.60%	98.00%	59.00%	63.91%	28.54%	87.00%	29.20%	3.028
Claude 3.7 Sonnet-20250219	61.82%	53.00%	93.40%	68.00%	67.14%	63.93%	85.60%	10.40%	4.768
Claude 4 Sonnet-20250514	93.13%	50.60%	98.60%	62.00%	66.94%	84.25%	96.20%	13.80%	4.638
Claude 4 Opus-20250514	94.75%	44.80%	88.60%	74.60%	65.73%	84.70%	99.40%	7.40%	4.602
Claude 3.7 Sonnet-20250219-thinking	94.14%	54.80%	98.00%	53.40%	67.74%	85.39%	63.60%	3.60%	4.806
DeepSeek-V3-250324	95.56%	48.80%	96.80%	83.60%	67.34%	77.63%	55.00%	3.60%	4.508
DeepSeek-R1-0528	95.15%	51.80%	97.40%	60.00%	66.53%	87.21%	46.80%	4.20%	4.608
DeepSeek-R1-Distill-Qwen-7B	81.21%	54.70%	93.52%	68.21%	54.44%	58.22%	22.20%	80.80%	3.222
Deepseek-R1-Distill-Llama-70B	91.90%	51.80%	58.60%	58.60%	64.31%	74.21%	7.00%	50.20%	2.570
DeepSeek-R1-Distill-Qwen-32B	89.70%	54.20%	95.60%	72.60%	59.68%	73.97%	12.20%	4.80%	3.830
Qwen3-235B-A22B	92.73%	47.60%	97.20%	54.60%	67.34%	84.02%	20.00%	18.60%	4.636
Qwen2.5-72B-Instruct	28.28%	45.40%	98.20%	55.80%	62.10%	52.51%	90.60%	14.20%	3.994
Qwen/QwQ-32B	93.54%	54.00%	95.80%	56.80%	60.69%	80.59%	15.80%	10.80%	3.222
Grok4-0709	77.78%	65.20%	96.60%	80.20%	65.12%	71.69%	80.00%	42.60%	4.160
Grok3	70.51%	52.20%	97.60%	80.80%	65.73%	66.44%	76.60%	26.00%	4.766

Model	SECQUE			EDTSUM			AT_T_Algorithm		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
GPT-4o-2024-05-13	31.05%	14.50%	19.46%	24.76%	11.31%	19.06%	47.25%	19.52%	26.34%
GPT-4.1-2025-04-14	24.59%	10.59%	14.71%	23.81%	9.91%	17.77%	52.90%	21.46%	28.87%
GPT-5-2025-08-07	29.54%	10.18%	16.33%	31.81%	13.57%	25.03%	46.39%	16.11%	27.67%
Gemini-2.5-Flash-thinking	33.60%	14.27%	21.03%	26.86%	11.83%	20.46%	52.57%	23.80%	34.26%
Gemini-2.5-Pro-thinking	35.01%	14.52%	21.19%	36.83%	17.25%	29.52%	51.57%	21.45%	30.29%
Llama3-8B	37.60%	15.79%	23.47%	22.49%	10.26%	17.47%	50.36%	18.40%	26.90%
Llama-3-70B-Instruct	34.78%	14.93%	21.36%	30.09%	14.44%	24.48%	51.50%	19.50%	27.60%
Llama-3.3-70B-Instruct	32.86%	14.41%	20.32%	35.16%	17.09%	28.72%	47.27%	17.97%	25.54%
Llama-4-Scout	39.00%	16.85%	24.11%	22.15%	10.33%	17.31%	47.33%	18.13%	27.80%
Claude 3.7 Sonnet-20250219	33.15%	13.34%	19.22%	29.14%	13.49%	23.42%	51.09%	21.85%	29.40%
Claude 4 Sonnet-20250514	32.72%	12.36%	18.81%	20.45%	8.66%	15.54%	52.05%	20.53%	29.46%
Claude 4 Opus-20250514	35.08%	13.70%	20.20%	40.17%	18.79%	33.73%	52.92%	21.22%	29.13%
Claude 3.7 Sonnet-20250219-thinking	34.26%	13.18%	19.61%	36.27%	17.02%	30.21%	50.90%	20.17%	28.47%
DeepSeek-V3-250324	29.90%	10.24%	16.45%	21.66%	8.81%	16.13%	51.44%	19.77%	27.94%
DeepSeek-R1-0528	27.59%	9.70%	15.42%	17.19%	7.06%	12.71%	50.49%	18.33%	26.43%
DeepSeek-R1-Distill-Qwen-7B	32.34%	11.51%	18.23%	20.37%	8.36%	15.43%	47.32%	16.81%	24.57%
Deepseek-R1-Distill-Llama-70B	33.95%	11.76%	19.66%	29.83%	13.34%	23.60%	42.10%	15.11%	23.51%
DeepSeek-R1-Distill-Qwen-32B	33.92%	13.62%	19.98%	29.19%	13.61%	23.57%	50.84%	19.42%	26.91%
Qwen3-235B-A22B	30.04%	11.42%	17.00%	18.64%	7.37%	13.98%	20.25%	9.17%	11.78%
Qwen2.5-72B-Instruct	23.13%	10.93%	14.47%	21.25%	9.14%	15.96%	45.45%	19.31%	25.77%
Qwen/QwQ-32B	14.05%	6.28%	8.77%	6.02%	2.62%	4.83%	21.12%	9.66%	12.37%
Grok4-0709	31.15%	12.94%	18.01%	23.93%	10.81%	18.76%	48.35%	19.60%	27.50%
Grok3	25.49%	11.65%	15.60%	24.00%	10.88%	18.77%	47.33%	20.22%	26.57%

Model	FNXL	FinQA	TATQA	FinRED	SD_T	SD_C	PM_A	PM_C	AT_T_Strategy
	GPT Eval (Label)				GPT Eval (Score)				
GPT-4o-2024-05-13	28.00%	52.00%	90.40%	33.80%	3.547	3.994	4.584	4.762	4.852
GPT-4.1-2025-04-14	16.40%	59.20%	91.20%	23.30%	4.330	4.325	4.765	4.822	4.955
GPT-5-2025-08-07	23.00%	71.80%	92.40%	18.40%	4.660	4.692	4.876	4.880	4.943
Gemini-2.5-Flash-thinking	19.20%	68.20%	92.40%	20.30%	4.032	3.903	4.794	4.872	4.909
Gemini-2.5-Pro-thinking	6.00%	65.60%	88.20%	18.90%	3.764	3.873	4.813	4.836	4.841
Llama3-8B	1.30%	32.80%	77.40%	33.10%	3.119	2.790	2.765	3.946	4.273
Llama-3-70B-Instruct	6.60%	48.60%	87.20%	24.00%	3.447	3.440	4.289	4.732	4.670
Llama-3.3-70B-Instruct	2.50%	47.40%	85.40%	19.10%	3.513	3.544	4.270	4.800	4.795
Llama-4-Scout	7.50%	58.40%	88.60%	17.20%	2.937	2.573	3.000	3.918	3.182
Claude 3.7 Sonnet-20250219	27.40%	64.40%	90.20%	19.60%	4.053	4.135	4.724	4.926	4.932
Claude 4 Sonnet-20250514	33.00%	74.20%	92.20%	17.90%	3.846	3.696	4.759	4.906	4.841
Claude 4 Opus-20250514	46.20%	69.20%	91.40%	21.60%	3.884	3.829	4.794	4.880	4.898
Claude 3.7 Sonnet-20250219-thinking	50.00%	70.00%	89.80%	21.10%	4.151	4.429	4.759	4.912	4.886
DeepSeek-V3-250324	9.70%	63.20%	88.60%	14.20%	3.805	3.569	4.778	4.780	4.864
DeepSeek-R1-0528	8.80%	61.40%	90.40%	19.10%	3.796	4.026	4.835	4.822	4.875
DeepSeek-R1-Distill-Qwen-7B	0.30%	47.60%	74.60%	19.90%	2.731	2.879	4.317	4.206	4.238
Deepseek-R1-Distill-Llama-70B	3.50%	60.00%	75.60%	19.90%	3.050	3.604	3.443	2.224	2.080
DeepSeek-R1-Distill-Qwen-32B	4.10%	63.60%	88.00%	18.90%	3.440	3.774	4.635	4.698	4.693
Qwen3-235B-A22B	23.30%	63.80%	89.60%	19.60%	3.173	4.119	4.787	4.812	4.500
Qwen2.5-72B-Instruct	17.90%	49.40%	88.60%	23.80%	3.484	3.679	4.546	4.834	4.818
Qwen/QwQ-32B	5.30%	59.60%	88.40%	19.10%	2.969	3.085	4.527	4.408	4.011
Grok4-0709	29.90%	61.20%	87.20%	21.80%	3.503	3.903	4.743	4.780	4.716
Grok3	15.10%	51.80%	85.00%	19.60%	3.840	3.800	4.762	4.872	4.932

Table 16: Complete evaluation results for application-based tasks in BizCompass.

I LLM Eval Prompts

Following G-Eval's setting (Liu et al., 2023), we select GPT-4o as the evaluation model.

LLM Eval Prompt for general QA

You are an expert evaluator. Your task is to score the candidate's answer based on both the correctness of the final answer and the logical soundness of its reasoning.

IMPORTANT INSTRUCTION

If the question contains multiple sub-questions (*e.g.*, 1., 2(a).), you MUST score each sub-question INDEPENDENTLY.

Evaluate the [Conclusion Correctness] and [Reasoning Soundness] for each sub-question in the candidate's answer, and strictly apply the following criteria to score EACH sub-question:

- 5 (Completely Correct): - Evaluation Points: The conclusion is correct, concepts are applied precisely, and the reasoning is clear, rigorous, and flawless.
- 4 (Incorrect Conclusion, Sound Reasoning - Minor Deviation): - Evaluation Points: The overall framework and method of reasoning are correct, but the conclusion is flawed due to minor calculation errors, data misinterpretations, or slight misunderstandings of boundary conditions. The core logic is sound.
- 3 (Incorrect Conclusion, Sound Reasoning - Major Deviation): - Evaluation Points: The general direction of the reasoning is correct, but a key assumption, formula application, or core step contains a significant error, leading to a major deviation in the conclusion.
- 2 (Correct Conclusion, Flawed Reasoning - Minor Flaws): - Evaluation Points: The conclusion happens to be correct, but the reasoning is not rigorous, showing minor flaws such as logical leaps, insufficient justification, or confusion of secondary concepts.
- 1 (Correct Conclusion, Flawed Reasoning - Severe Flaws): - Evaluation Points: The conclusion happens to be correct, but the reasoning has fundamental errors, such as reversed causality or the use of a completely wrong model/theory. The connection between the reasoning and the conclusion is purely coincidental.
- 0 (Completely Incorrect): - Evaluation Points: Both the conclusion and the reasoning process have severe errors, or the answer is completely off-topic and provides little to no valid reasoning.

OUTPUT FORMAT:

Return ONLY a JSON object. The object must contain two fields: 1. 'score': The arithmetic mean of all sub-question scores. 2. 'scores_per_question': An object containing the independent score for each sub-question. The format must be exactly as follows:

{
"score": <arithmetic mean score>,
"scores_per_question": {
"<question_number_1>": <integer from 0 to 5>,
"<question_number_2>": <integer from 0 to 5> }
}

J More Results

J.1 Statistical Analysis of Business Scenarioss

To explore the potential applications of LLMs in business, we begin by defining three representative business scenarios—analysis, trading, and consulting—based on authoritative external sources and expert consultation. Specifically, analysis follows the definition provided in the International Labour Organization's ISCO-08 classification (Ganzeboom, 2010), trading is defined according to the description published by Indeed UK (Kaniel and Liu, 2006), and consulting is defined using the official description provided by LinkedIn Talent Solutions (Carliner et al., 2015).

Definition 1 (Analysis / Business Analysis)**.** *Analysis involves assisting clients in evaluating information that influences investment programmes in both public and private institutions. Professionals in this role are commonly referred to as financial analysts.*

Definition 2 (Trading)**.** *Trading involves buying and selling stocks and shares in financial markets with other participants who hold an interest in the stock market and finance.*

Definition 3 (Consulting)**.** *Consulting involves providing specialized expertise and knowledge—either independently or through a consulting firm—to help businesses achieve goals and solve problems.*

In order to validate the above definitions, we conduct a systematic analysis of O*NET. The O*NET database is organized in a hierarchical structure: at the top level are career clusters, which group related occupational domains (e.g., financial services). Each cluster is further divided into sub-clusters that capture more specific functional areas (e.g., financial strategy & investments). At the lowest level are occupations, each documented with a job title, a definition, and a set of task descriptions. Building on this structure, our analysis proceeds in three steps:

Step One: Cluster Selection and Occupation Collection. We begin by selecting three career clusters most relevant to business practice—management & entrepreneurship, marketing & sales, and financial services. Together, these clusters comprise 14 unique sub-clusters and 156 occupations in total. Of these, 126 occupations fall within the scope of our three scenario definitions. The distribution of in-scope occupations is relatively balanced, with 59 from management & entrepreneurship, 41 from financial services, and 26 from marketing & sales.

Step Two: Occupation-to-Scenario Mapping. Each occupation is assigned to one of the three business scenarios by matching its related information against definitions 1 to 3. The classification is performed by an LLM, with the mapping guided by the following prompt:

Step Three: Scenario Coverage Calculation. In the final step, we merge the occupations collected and remove duplicates, thereby yielding a unified occupation set $\tilde{J}$. For each occupation $j \in \tilde{J}$, we record its assigned category $c_j \in \{\text{consulting}, \text{analysis}, \text{trading}, \text{none}\}$, along with two associated quantities: the number of employed persons $e_j \geq 0$ and the projected number of job openings $o_j \geq 0$. Based on this setup, our interest lies in the relative importance of the three core business scenarios $C = \{\text{consulting}, \text{analysis}, \text{trading}\}$. To capture this, we define two proportional measures. The first is the employment share, which quantifies the proportion of total employment represented by a given category:

$$S^{(\text{emp})}(c) = \frac{\sum_{j \in \tilde{J}} e_j \, 1\{c_j = c\}}{\sum_{j \in \tilde{J}} e_j} \times 100\%, \quad c \in C,$$

where the denominator aggregates employment across all categories (including none), while the numerator sums only employment in category c. Here, $1\{\cdot\}$ denotes the indicator function. The second is the job-opening share, which quantifies the proportion of projected job openings represented by a given category:

$$S^{(\text{open})}(c) = \frac{\sum_{j \in \tilde{J}} o_j \, 1\{c_j = c\}}{\sum_{j \in \tilde{J}} o_j} \times 100\%, \quad c \in C,$$

where the denominator aggregates openings across all categories (including none), while the numerator sums only openings in category c. The employment share reflects the current labor market distribution, whereas the job-opening share captures future demand.

Metric	Consulting	Analysis	Trading	Total
$S^{(\text{emp})}(c)$ (%)	38.76	33.60	23.20	95.56
$S^{(\text{open})}(c)$ (%)	35.01	30.72	29.52	95.25

Table 17: Employment and job-opening shares of three representative business scenarios.

The statistical results are summarized in Table 17, which demonstrates that the three representative business scenarios together cover the vast majority of both employment and job openings.

J.2 Task Coverage Validation

We now assess whether the three roles, Analysis, Trading, and Consulting, align with the observed work content. Using O*NET, we collect all tasks from the 126 in-scope occupations and, for each

Analysis		Consulting		Trading	
Task	Coverage (%)	Task	Coverage (%)	Task	Coverage (%)
asset pricing	36.52	market trend analysis	31.05	asset pricing	30.40
market trend analysis	44.76	risk management	74.55	market trend analysis	30.09
risk management	76.12	strategy development	66.84	strategy development	39.51
portfolio management	43.91	portfolio management	29.14	algorithmic trading	12.16
fraud detection	48.41	fraud detection	25.77		
financial document analysis	55.90	financial document analysis	31.57		
Overall	88.20	Overall	91.13	Overall	70.52

Table 18: Task coverage by role.

Note. Tasks may match multiple representative tasks; *Overall* is not a sum of rows but the share of tasks matched at least once within each role.

role, check whether each task matches that role's representative tasks. Our goal is to measure how much of the work of each role is captured by its representative set and to ensure the internal consistency of the design.

Step One: Collect and summarize tasks. Starting from the 126 in-scope occupations, we extract all tasks from O*NET and deduplicate them. For each role $c \in \{\text{Analysis}, \text{Trading}, \text{Consulting}\}$, let K_c be the set of unique tasks associated with c, and let $T_c = |K_c|$. In our data: $T_{\text{Analysis}} = 1068$, $T_{\text{Trading}} = 329$, and $T_{\text{Consulting}} = 947$.

Step Two: Test alignment with real tasks. We evaluate a set of 16 task–role pairs. We assign the following subsets to each role:

- **Analysis**: *asset pricing*, *market trend analysis*, *risk management*, *portfolio management*, *fraud detection*, *financial document analysis*.
- **Trading**: *asset pricing*, *market trend analysis*, *strategy development*, *algorithmic trading*.
- **Consulting**: *market trend analysis*, *risk management*, *strategy development*, *portfolio management*, *fraud detection*, *financial document analysis*.

Let S_c denote the subset for role c. For each task $k \in K_c$ and each representative task $\tau \in S_c$, we use an LLM (*LLM*, temperature $= 0$) to decide whether k is related to τ under a fixed yes/no rule. The decision is recorded as

$$f_{\tau,c}(k) \in \{0, 1\} \quad (1 = \text{relevant}, \ 0 = \text{not relevant}).$$

Step Three: Task Coverage Calculation. In this step, we assess how completely the representative tasks account for the collected tasks within each role. Let K_c, T_c, and S_c be as defined above, and let $f_{\tau,c}(k) \in \{0, 1\}$ denote the relevance decision.

Based on this setup, we define two measures. The first is the share of each representative task within its role:

$$H_{\tau,c} = \frac{1}{T_c} \sum_{k \in K_c} f_{\tau,c}(k) \times 100\%.$$

This reports the percentage of tasks in role c that are covered by τ.

The second is the overall coverage of a role. A task counts as covered if it matches at least one representative task:

$$U_c = \{\, k \in K_c : \exists\, \tau \in S_c \text{ with } f_{\tau,c}(k) = 1 \,\},$$

$$\text{Coverage}(c) = \frac{|U_c|}{T_c} \times 100\%.$$

Because a task may match more than one representative task, the shares $H_{\tau,c}$ do not sum to 100%. By contrast, $\text{Coverage}(c)$ counts each task once.

The statistical results are summarized in Table 18, which show that the representative task sets cover the large majority of observed tasks within each role. Within roles, individual representative tasks already account for substantial portions—for example, *risk management* covers 76.12% of Analysis tasks and 74.55% of Consulting; *market trend analysis* covers 44.76% in Analysis and 31.05% in Consulting; and in Trading the main contributors are *strategy development* (39.51%) and *asset pricing* (30.40%). Aggregating within roles, the *Overall* row reaches 88.20% for Analysis, 91.13% for Consulting, and 70.52% for Trading. Taken together, these results indicate broad coverage: most in-scope tasks can be matched to at least one representative task, suggesting that the benchmark reflects real work and is suitable for a wide range of practical applications.

Prompt for Statistic-GeneralQA

ROLE:
You are an expert researcher in statistics, specializing in statistical theory and methodology. You should provide a rigorous and complete answer to the question. The answer must be structured in a step-by-step format that directly corresponds to each sub-question from the input, maintaining the original numbering and lettering (e.g., 1., 2(a)., 2(b).).

OUTPUT FORMAT:
Return ONLY the answer as a single string. The string must be a complete, step-by-step solution. The answer must use the exact same notation and LaTeX expressions as provided in the question.

EXAMPLE

INPUT:
Background:
Research Question. How does the homoscedasticity assumption simplify variance in mixed-effects models, and what are its practical limitations?
Setting:
Response y decomposes into orthogonal effects: $y = \sum a_G$, with fixed factors J and random factors. Conditional variance is $var(y|J) = \sum_{G \notin J} E(a_G^2|J) + \sum_{G1 \neq G2 \notin J} E(a_{G1} a_{G2}|J)$.
Under homoscedasticity (F3), this simplifies to $var(y|J) = \sum_{G \notin J} \sigma_G^2$.
Questions:
1. Derive the simplified variance under (F3) and show it implies conditionally uncorrelated random effects.
2. Critique (F3) with a clinical trial example where it fails, and discuss consequences for inference if violated but the simplified model is used.

OUTPUT:
1. Under (F3), $E(a_G^2|J) = \sigma_G^2$ (constant) for random G, and $E(a_{G1} a_{G2}|J) =$ constant for $G1 \neq G2$. Since unconditional orthogonality implies $E(a_{G1} a_{G2}) = 0$, the constant must be 0, so conditional covariances vanish. Thus, $var(y|J) = \sum_{G \notin J} \sigma_G^2$, with effects conditionally uncorrelated.
2. In a trial with fixed factor Treatment (Placebo, Low Dose, High Dose) and random factor Patient, high-dose responses may vary more due to side effects, violating constant σ_R^2 across treatments. If violated but simplified model used, variance estimates are wrong, leading to invalid p-values, confidence intervals, and hypothesis tests for treatment effects.

Prompt for Economics-Single Choice

ROLE:
You are an expert economic researcher specializing in econometrics and empirical economics. You should carefully analyze the given question and options, determine the single correct choice, and output only the letter of the correct option.

OUTPUT FORMAT:
Return ONLY the answer as a single string containing the chosen letter. The string must not include any other text or explanations.

EXAMPLE

INPUT:
Background:
In ordinary least squares (OLS) regression, a key assumption is homoskedasticity: the error terms have constant variance across all observations. If violated (heteroskedasticity), standard errors may be biased, leading to invalid inference.
Question:
In which of the following scenarios is heteroskedasticity most likely to occur in a regression of income on education?
Options:
A) Errors are randomly distributed and independent of education level.
B) All individuals report income accurately, with no measurement error.
C) Variance in income is larger for highly educated individuals due to diverse job opportunities.
D) The sample size is small, but errors are normally distributed.

OUTPUT:
C

Prompt for Finance-Multiple Choice

ROLE:
You are an expert financial researcher specializing in financial theory and quantitative methods. You should carefully analyze the given question and options, determine all correct choices, and output only the letters of the correct options separated by commas (e.g., “A,B").
OUTPUT FORMAT:
Return ONLY the answer as a string containing the chosen letters separated by commas (e.g., “A,B"). The string must not include any other text or explanations. If no options are correct, return an empty string.

EXAMPLE
INPUT:
Background:
The Capital Asset Pricing Model (CAPM) describes the relationship between systematic risk and expected return for assets, assuming investors are rational and markets are efficient.
Question:
Select all assumptions of the CAPM that are essential for its derivation.
Options:
A: Investors have homogeneous expectations about asset returns.
B: There are no taxes or transaction costs.
C: All investors can borrow and lend at the risk-free rate.
D: Assets are infinitely divisible.

OUTPUT:
A,B,C,D

Prompt for OM-TableQA

ROLE:
You are an expert researcher in Operations Management, specializing in mathematical modeling and optimization. You should provide a rigorous and complete answer to the question. The answer must be structured in a step-by-step format that directly corresponds to each sub-question from the input, maintaining the original numbering and lettering (e.g., 1., 2(a)., 2(b).). Your entire analysis must be strictly confined to the provided background and table information.

OUTPUT FORMAT:
Return ONLY the answer as a single string. The string must be a complete, step-by-step solution. The answer must use the exact same notation and LaTeX expressions as provided in the question. In your reasoning, you must explicitly cite the tables the data is from (e.g., "according to Table 1...") and any formulas defined in the question (e.g., "using eq. (1)...").

EXAMPLE
INPUT:
Background:
Research Question. A manufacturing company wants to optimize its inventory management for a key component. The goal is to determine the optimal order quantity that minimizes the total annual inventory cost, which consists of ordering costs and holding costs.
Data / Model Specification:

Table 1: Inventory Parameters for Component XJ-100

Parameter	Symbol	Value
Annual Demand	D	10,000 units
Ordering Cost per Order	S	$50
Annual Holding Cost per Unit	H	$4

The Economic Order Quantity (EOQ) is calculated using the following formula:
$Q^{\wedge} = \sqrt{\frac{2DS}{H}}$ (1)
The Total Annual Cost (TC) for any given order quantity Q is:

$$TC(Q) = \frac{D}{Q}S + \frac{Q}{2}H \quad (2)$$

Questions:
1. Using the data from Table 1 and eq. (1), calculate the Economic Order Quantity $Q^{\wedge}$ for component XJ-100.
2. The current policy is to order 1,000 units at a time. Using eq. (2), calculate the total annual cost for the current policy ($Q = 1000$) and for the optimal EOQ policy ($Q = Q^{\wedge}$) you found in part 1. How much money does the company save annually by switching to the EOQ model?

OUTPUT:
1. To calculate the Economic Order Quantity ($Q^{\wedge}$), we substitute the parameter values from Table 1 into eq. (1):
Annual Demand, $D = 10,000$ units.
Ordering Cost, $S = \$50$ per order.
Holding Cost, $H = \$4$ per unit per year.
Using eq. (1):

$$Q\hat{} = \sqrt{\frac{2 \times 10000 \times 50}{4}} = \sqrt{\frac{1,000,000}{4}} = \sqrt{250,000} = 500$$

The Economic Order Quantity is 500 units.

2. To find the cost savings, we first calculate the total annual cost for both policies using eq. (2) and the data from Table 1.

Cost of Current Policy (Q = 1,000):

$$TC(1000) = \frac{10000}{1000}(50) + \frac{1000}{2}(4) = 10(50) + 500(4) = 500 + 2000 = \$2,500$$

Cost of EOQ Policy (Q = 500):

$$TC(500) = \frac{10000}{500}(50) + \frac{500}{2}(4) = 20(50) + 250(4) = 1000 + 1000 = \$2,000$$

The total annual cost is minimized at the EOQ, where the annual ordering cost equals the annual holding cost.

Annual Savings: The savings from switching policies is the difference between the two costs: Savings = $TC(1000) - TC(500) = \$2,500 - \$2,000 = \$500$

The company saves $500 annually by switching to the EOQ model.

Prompt for M&A

ROLE:

You are an expert trader specializing in market trend analysis. In this task, you will be given Mergers and Acquisitions (M&A) news articles or tweets. Your task is to classify each article or tweet based on whether the mentioned deal was completed or remained a rumour. Your response should be a single word - either 'complete' or 'rumour' - representing the outcome of the deal mentioned in the provided text.

EXAMPLE

INPUT:

Microsoft announced today the successful completion of its $68.7 billion acquisition of Activision Blizzard. The deal, which received final regulatory approval last week, makes Microsoft the world's third-largest gaming company by revenue.

OUTPUT:

complete

Prompt for Taiwan

ROLE:

You are an expert consultant specializing in fraud detection. You should predict whether the company will face bankruptcy based on the financial profile attributes provided in the following text. Respond with only 'no' or 'yes', and do not provide any additional information.

EXAMPLE

INPUT:

The client has attributes: Bankrupt: 0.450, ROA(C) before interest and depreciation before interest: 0.120, ROA(A) before interest and after tax: 0.089, Operating Gross Margin: 0.234, Realized Sales

Gross Margin: 0.567, Operating Profit Rate: 0.156, Current Ratio: 0.890, Quick Ratio: 0.723, Debt ratio%: 0.834, Net worth/Assets: 0.166, Total Asset Turnover: 0.445, Cash Flow to Total Assets: 0.078, Net Income to Total Assets: 0.045, Interest Coverage Ratio: 2.340.

OUTPUT:
no

Prompt for FinQA

ROLE:
You are an expert analyst specializing in market trend analysis. Given the financial data and expert analysis, please answer this question:

EXAMPLE
INPUT:
The company's revenue for 2021 was $500 million and for 2022 was $600 million. Operating expenses were $300 million in 2021 and $350 million in 2022. What was the percentage increase in revenue from 2021 to 2022?

OUTPUT:
0.2

Prompt for AT-T-text

ROLE:
You are an expert algorithmic trading strategist. You specialize in articulating complex quantitative frameworks and methodologies into clear, high-level strategy descriptions. Your task is to provide a comprehensive, descriptive answer that synthesizes the provided information into a coherent strategic process. The answer must be structured in a step-by-step format that directly corresponds to each sub-question from the input, maintaining the original numbering (e.g., 1., 2., 3.). Your entire analysis and the framework you create must be strictly and exclusively based on the information and data points provided in the text.

EXAMPLE
INPUT:
Framework for a Mean-Reversion System
A mean-reversion strategy is based on the statistical premise that asset prices tend to revert to their long-term average. This framework uses Bollinger Bands to identify trading opportunities.
Key Components
[Definition] Moving Average (MA): The average price of an asset over a specified period (e.g., 20 days).
[Definition] Bollinger Bands: A channel plotted two standard deviations above and below the moving average. Prices are considered high when above the upper band and low when below the lower band.
Three-Step Strategic Process
1. Establish Price Channel: Continuously calculate the 20-day simple moving average and the upper/lower Bollinger Bands for the target asset.
2. Generate Entry Signals: A buy signal is generated when the asset's price closes below the lower Bollinger Band. A sell (short) signal is generated when the price closes above the upper Bollinger

Band.
3. Define Exit Rule: The position is closed (exited) when the asset's price reverts and crosses back over the 20-day moving average.
Describe a complete, three-step strategy for a mean-reversion trading system based on the provided framework.

OUTPUT:
The strategy for this mean-reversion system is a three-step process designed to capitalize on temporary price deviations from a central average. First, the system establishes a dynamic price channel by continuously calculating a 20-day moving average and its corresponding upper and lower Bollinger Bands, which are set at two standard deviations from the average. Second, entry signals are generated based on price breaches of this channel: a buy signal is triggered when the price closes below the lower band, indicating an oversold condition, while a sell signal is triggered on a close above the upper band, indicating an overbought condition. Finally, the exit rule for any open position is explicitly defined by the price reverting to its historical mean, with the position being closed as soon as the price crosses back over the 20-day moving average.

Prompt for PM-A

ROLE:
You are a portfolio management analyst with a strong background in quantitative finance and asset pricing theory. You must provide a rigorous and complete answer to the question. The answer is sting and must be structured in a clear, step-by-step format that shows all calculations. Your entire analysis must be strictly confined to the provided background, formulas, and data.

EXAMPLE

INPUT:
You are a portfolio management analyst tasked with evaluating the expected return of a potential new stock for your fund, 'Innovate Corp'. To determine if the stock's potential return justifies its risk, you will use the Capital Asset Pricing Model (CAPM), a cornerstone of modern financial theory. The Capital Asset Pricing Model (CAPM) describes the relationship between systematic risk and expected return for assets. The formula is given in Formula 1:

$$E[R_i] = R_f + \beta_i(E[R_m] - R_f)$$

Where: $E[R_i]$ is the expected return of the investment. R_f is the risk-free rate. β_i (the beta of the investment) is the measure of the asset's systematic risk. $E[R_m]$ is the expected return of the market.
Your team has provided you with the following data: * Risk-Free Rate (R_f): 4%, * Expected Market Return ($E[R_m]$): 10%, * Innovate Corp. Stock Beta (β_i): 1.2. Using the Capital Asset Pricing Model (Formula 1), what is the required rate of return for Innovate Corp. stock?

OUTPUT:
Answer: The goal is to calculate the expected return of the stock using the CAPM formula.
Step 1: Identify the given parameters in decimal format. * Risk-Free Rate, $R_f = 0.04$. * Expected Market Return, $E[R_m] = 0.10$. * Stock Beta, $\beta_i = 1.2$.

Step 2: Substitute the values into the CAPM formula. Using Formula 1:

$$E[R_i] = 0.04 + 1.2 \times (0.10 - 0.04)$$

Step 3: Calculate the result. First, calculate the market risk premium:

$$E[R_m] - R_f = 0.10 - 0.04 = 0.06$$

Now, multiply by beta:

$$1.2 \times 0.06 = 0.072$$

Finally, add the risk-free rate:

$$E[R_i] = 0.04 + 0.072 = 0.112$$

*Conclusion: The required rate of return for Innovate Corp. stock, according to the CAPM, is 11.2%.

Prompt for RM-C

ROLE:
You are an expert risk management consultant. You specialize in translating qualitative risk assessments into formal, quantitative, and actionable risk management frameworks. You must provide a detailed and structured policy or framework that directly answers the questions. The answer must be structured in a step-by-step format that directly corresponds to each sub-question from the input, maintaining the original numbering (e.g., 1., 2., 3.). Your entire analysis and the framework you create must be strictly and exclusively based on the information and data points provided in the text. You must must explicitly cite the tables the data is from (e.g., "according to Table 1...") and any formulas defined in the question (e.g., "using eq. (1)...").

EXAMPLE
INPUT:
Our firm's positive thesis on Innovate Inc. is predicated on the continued high-margin dominance of its 'Pro-Suite' software. This single product line is the company's primary profit engine.
[Internal Risk Assessment]
Innovate Inc.'s primary vulnerability is its over-reliance on the Pro-Suite, which accounts for 60% of revenue and an estimated 80% of operating profit. A new competitor, Disruptor Corp., has recently launched a competing product at a 40% lower price point. Our base-case forecast assumes Innovate Inc. will lose 10% market share over the next two years but will maintain its pricing. However, the report highlights a more aggressive risk scenario where Disruptor Corp. captures 30% of the market within 18 months, forcing Innovate to cut prices by 20% to remain competitive. This would severely impact profitability.
Question:
1. Identify the single most critical risk to Innovate Inc.'s valuation based on the report.
2. Propose a key metric to monitor this risk.
3. Define a specific, quantitative trigger for this metric that would require a mandatory portfolio action, and state what that action should be.

OUTPUT:

1. The most critical risk is the severe margin and market share compression for its 'Pro-Suite' software, which is the company's primary profit driver, accounting for an estimated '80% of operating profit'. The threat comes from a new competitor offering a significantly lower-priced alternative.
2. The key metric to monitor this risk will be a combination of Pro-Suite's Average Selling Price (ASP) and its quarterly market share, as reported in the company's financial statements and industry analysis reports.
3. A mandatory portfolio action will be triggered if either of the following conditions is met in a single quarter: a) Pro-Suite's reported market share drops by more than 5% sequentially, or b) its reported ASP decreases by more than 7% sequentially. If this trigger is activated, the mandatory action is an immediate 25% reduction of our position in Innovate Inc. and a formal re-underwriting of the entire investment thesis based on the new competitive reality.