Strategic Business Management in the Banking Sector: Enhancing Operational Efficiency and Growth

Vidhi Mardia

Department of Finance, Trine University, Indiana, USA
vidhimardia@gmail.com

Abstract. Effective business management is a cornerstone of success in the banking sector, influencing financial performance, customer satisfaction, and long-term sustainability. This study examines key management strategies in banking, focusing on operational efficiency, leadership approaches, and customer relationship management. By analyzing market dynamics, regulatory considerations, and financial decision-making processes, this research highlights best practices that contribute to sustainable growth. Emphasis is placed on optimizing banking operations through structured management frameworks, workforce development, and customer-centric strategies. The findings suggest that a strong business management foundation enables banks to navigate economic fluctuations, strengthen competitiveness, and foster customer trust, ensuring resilience in an evolving financial landscape.

Keywords: Banking sector, Strategic management, Operational efficiency, Growth, Customer relationship management

1 Introduction

Modern enterprises rely on extensive datasets to extract meaningful patterns, identify trends, and support strategic decision-making [34]. The data anal- ysis workflow typically commences with defining broad objectives, followed by an iterative process of formulating queries, extracting insights, and synthesiz- ing results into actionable recommendations. Figure 1 illustrates this iterative analytical framework.

Traditional analytical methodologies predominantly utilize tools such as Jupyter notebooks and interactive dashboards [6]. These approaches demand considerable manual effort and a strong foundation in data science and domain-specific expertise. Recent advancements in large language models (LLMs) have enabled AI-driven systems capable of assisting in various analytical tasks [36]. However, existing AI-based analytical frameworks, including Pandas Agent [20] and Code Interpreter [37], remain limited to executing isolated, predefined queries, such as computing statistical metrics for specific models.

A. Jain and S. Gupta (eds.), *Proceedings of the International Conference on Policies, Processes and Practices for Transforming Underdeveloped Economies into Developed Economies (PPP-UD 2025)*, Advances in Economics, Business and Management Research 357,
https://doi.org/10.2991/978-94-6463-894-3_3

A fundamental limitation of current AI-based data analytics tools stems from the nature of existing benchmarks, which predominantly assess single-step code execution rather than multi-step analytical reasoning [19]. The effective evaluation of LLM-driven analytical agents requires a structured assessment framework that considers context, reasoning, and output validity. Current evaluation methods often involve human judgment, leading to inconsistencies and potential biases. Approaches such as InsightPilot [26] and InsightLens [1] have introduced AI agents for comprehensive data analytics; however, these rely on labor-intensive human evaluations.

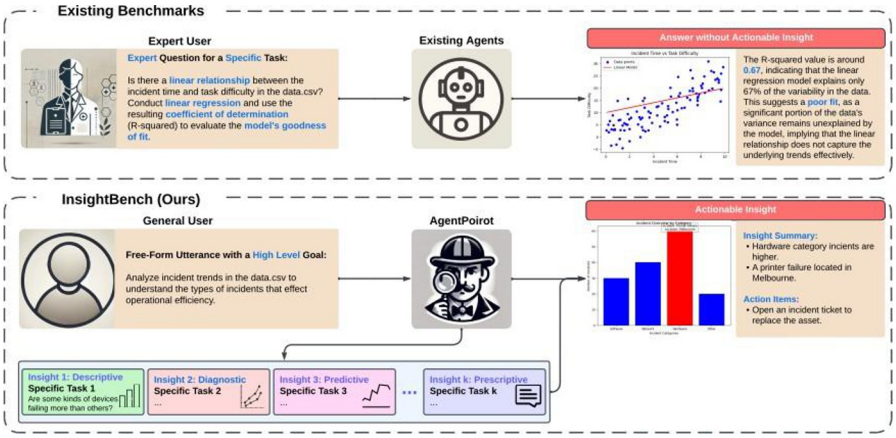

Fig. 1: Traditional benchmarks (top) focus on assessing agents based on predefined analytical queries with structured outputs, often requiring domain expertise for query formulation. In contrast, InsightBench (bottom) evaluates LLM-based agents on an end-to-end data analytics workflow, measuring their ability to interpret user queries, autonomously select analytical tasks, extract multidimensional insights, and generate concise, context-aware recommendations, as exemplified by AgentPoirot.

To standardize and enhance the evaluation of AI-driven analytical agents, a structured benchmarking framework is necessary. This framework should address two critical challenges: identifying the most relevant investigative queries within a dataset and ensuring that generated insights align with expert-driven interpretations. Given the vast range of potential questions and interpretations, a well-defined evaluation approach requires datasets with explicit analytical objectives, comparable to those established by experienced analysts.

To address this gap, InsightBench has been developed as a benchmark consisting of 100 datasets derived from business workflow data available on the ServiceNow platform [40]. These datasets span diverse categories, including Incident Management, User Records, Financial Transactions, Inventory Manage-

ment, and Enterprise Goal Tracking. The structured nature of these datasets facilitates the evaluation of AI-driven analytical agents in enterprise decision-making scenarios.

Each dataset has been carefully curated by expert annotators to define explicit objectives and embed meaningful insights. These insights are paired with guiding questions, associated code implementations, extracted values, and visual representations, culminating in summarized findings and actionable recommendations. This structured approach ensures that AI-based analytical models generate insights that align closely with human expert assessments.

A robust evaluation mechanism is essential to quantify the alignment between generated insights and reference insights. Inspired by G-Eval [23], which utilizes GPT-4 for qualitative assessment, an alternative evaluation method, LLaMA-3-Eval, is introduced. LLaMA-3-Eval leverages the open-source LLaMA-3 model [42] to provide a cost-effective alternative while maintaining high consistency with G-Eval in ranking analytical performance.

InsightBench serves as a foundation for evaluating various LLM-driven analytical agents, including a modified Pandas Agent capable of multi-step data exploration. Additionally, a baseline model, AgentPoirot, is introduced to enhance insight extraction through structured prompts aligned with Descriptive ("What happened?"), Diagnostic ("Why did it happen?"), Predictive ("What is likely to happen?"), and Prescriptive ("What actions should be taken?") analytics. Experimental evaluations demonstrate that AgentPoirot achieves higher LLaMA-3-Eval scores compared to existing analytical agents, including Pandas Agent.

This study presents the following key contributions:

- Development of InsightBench, a benchmark designed to evaluate LLM-driven agents on multi-step, end-to-end data analytics.
- Comparative assessment of various open-source and proprietary models, highlighting the superior performance of AgentPoirot over existing analytical frameworks, including Pandas Agent.
- Validation of LLaMA-3-Eval as an effective alternative to G-Eval for evaluating AI-generated insights.

The novelty of this research lies in explicitly linking strategic management principles with operational efficiency models tailored for the banking sector, integrating customer-centric growth frameworks with regulatory adaptability—an approach not extensively explored in prior studies.

2 Related Works

The development of music recommender systems has been a central theme in information retrieval and user personalization. Wira [45] examined collaborative filtering algorithms on the Million Song Dataset, highlighting the competitiveness of simple popularity-based models compared to probabilistic matrix factorization, with user-based and item-based collaborative filtering achieving superior

performance. Extending this perspective, Wira [44] traced the evolution of music consumption from live performances and physical media to digital streaming and playlist-based systems. The study emphasized the role of playlist generation strategies, including similarity-based and case-based approaches, while also introducing advanced techniques such as the Global-Local Similarity Function (GLSF). Furthermore, Wira [46] investigated deep neural networks for multitrack audio transformation, underscoring the broader applicability of AI across music production and intelligent sound design.

Parallel to these efforts, research has also expanded into adjacent computational and visualization domains. Thukkani [41] introduced the ioa++ framework, which advances distributed systems modeling with dynamic composition and concurrency-aware scheduling, demonstrating efficiency improvements for real-world communication protocols. Reddy [39] evaluated visualization methods for categorical data, revealing perceptual biases in parallel sets and hammock plots and proposing the common angle plot as a robust alternative.

In healthcare analytics, Arul [4] applied an Adaptive-Network-based Fuzzy Inference System (ANFIS) for diabetes risk prediction, achieving improved accuracy in chronic disease diagnostics. Idnani [14] proposed a correlation-driven framework for multivariate time series forecasting, demonstrating enhanced predictive accuracy across real-world datasets. In a related socio-economic study, Idnani [15] examined human capital depreciation and gender-specific wage disparities in Italy, offering insights into long-term labor market dynamics.

Emerging technologies have also shaped computational paradigms. Dhami [12] provided a comprehensive introduction to quantum machine learning, highlighting its applications across optimization and data science. Dhami [11] also analyzed the expansion of exchange-traded funds (ETFs) and their unintended consequences on volatility, liquidity, and efficiency in financial markets. Complementing this, Makhija [28] emphasized supply chain optimization in the U.S. through automation and intelligent infrastructure, while another work [27] proposed an automated image-based resistance assay for plant phenotyping.

In computer vision and workflow analysis, Malhotra [32] introduced a genetic algorithm-based image compression method using triangle representations, significantly reducing image size while maintaining fidelity. Malhotra [33] further proposed a hybrid strategy for predicting execution time in distributed workflows modeled as directed acyclic graphs. Expanding into the ethical dimension of AI, Malhotra [31] analyzed value assumptions in machine learning publications. Additional contributions include student placement optimization [29], and software quality assurance through hybrid automated and human testing [30].

Finally, wireless communication has gained prominence in Industry 4.0 and smart mobility systems. Katkar [16] presented a discrete event simulation approach for optimizing wireless-enabled on-demand transportation systems, while Katkar [17] investigated wireless technologies for industrial automation, comparing the trade-offs of 5G and LoRa protocols for efficiency and scalability.

Taken together, these works highlight the breadth of computational research across domains. Within this spectrum, music recommendation systems emerge

as a critical area that bridges human experience, AI, and large-scale data-driven personalization.

3 InsightBench – A Benchmark for Enterprise Data Analytics

InsightBench comprises datasets derived from the ServiceNow platform, a system extensively utilized for managing business workflows and operations. This platform maintains structured records across various enterprise functions, including incident tracking, user management, and asset monitoring. The benchmark incorporates 100 tabular datasets categorized into five thematic domains, as illustrated in Figure 3.

3.1 Benchmark Development Process

InsightBench serves as an evaluation framework for assessing the performance of LLM-driven agents in enterprise data analysis. The creation of this benchmark follows a structured four-step methodology: (1) extracting and defining schema elements from ServiceNow data tables (Section 2.1.1), (2) integrating structured trends to introduce analytical depth (Section 2.1.2), (3) generating synthetic entries aligned with embedded patterns (Section 2.1.3), and (4) compiling ground-truth analytical notebooks for standardized assessment (Section 2.1.4). The complete workflow is depicted in Figure 2. A comprehensive thematic breakdown of datasets is presented in Appendix B.3, while the evaluation procedure is detailed in Section 2.2.

3.2 Construction of InsightBench Datasets

Each dataset consists of 500 synthetically generated entries stored in CSV format, ensuring computational efficiency while maintaining enterprise-scale analytical relevance. To enhance realism, structured enterprise data is combined with synthetic insights, allowing meaningful evaluation of analytical models.

Data Tables The dataset design process begins by selecting key ServiceNow tables. For example, an incident table records operational disruptions, including hardware failures and software malfunctions. This table is structured with attributes such as timestamps, descriptions, and assigned personnel, facilitating in-depth data analysis (Figure 2(a)). Selected columns are refined to retain analytical significance while excluding extraneous information.

Embedding Analytical Patterns Statistical modeling techniques are employed to introduce structured trends and anomalies within dataset columns.

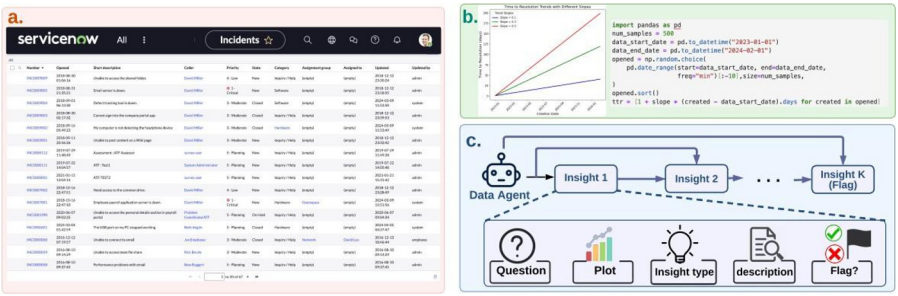

Fig. 2: Benchmark Development Stages: (a) Showcases a sample incidents table from ServiceNow under the Incident Management theme, detailing schema and data attributes. (b) Illustrates dataset construction by embedding a linear increasing trend in incident resolution times, highlighting the influence of the slope parameter on trend intensity. (c) Represents the multi-step annotation pipeline, including a guiding question, data visualization, extraction of descriptive insights, and categorization of insight types.

Regression-based approaches ensure consistency in embedded insights. For instance, an increasing trend in resolution times within a service management dataset follows the mathematical formulation:

$$TTR = 1 + \text{slope} \cdot (\text{incident open date} - \text{data start date}) \quad (1)$$

where slope determines the rate of increase (Figure 2(b)).

Categories of Insights The insights generated in InsightBench are classified into four key analytical types:

- **Descriptive**: These insights summarize past data to highlight patterns and distributions. Example: Visualization of incident categories over time.
- **Diagnostic**: These insights investigate causal relationships between variables. Example: Identification of common incident keywords through textual analysis.
- **Predictive**: These insights forecast future events based on historical data. Example: Time-series forecasting of incident volumes.
- **Prescriptive**: These insights propose actionable strategies derived from analytical findings. Example: Recommending interventions to mitigate rising incident rates.

Synthetic Data Generation A combination of two techniques ensures the authenticity of dataset attributes while preserving enterprise relevance. The entire pipeline is implemented in Python for scalability and reproducibility.

- **Random Sampling**: Categorical fields such as incident types, timestamps, and status indicators are populated using controlled random sampling from predefined value distributions.
- **Context-Aware Generation via LLMs**: Text-based fields, including incident descriptions and user feedback, are generated using LLMs such as GPT-4. These models produce structured, contextually appropriate entries aligned with enterprise data attributes.

Further implementation specifics are discussed in Appendix B.1, with dataset generation prompts outlined in Appendix D.

3.3 Ground-Truth Analytical Notebooks

To ensure objective evaluation, 100 expert-annotated Jupyter notebooks have been created, each corresponding to a dataset theme, such as service management or financial analytics. These notebooks adhere to a predefined structure (Figure 2(c)):

- **Dataset Overview and Analytical Objective**: Each notebook begins with a dataset description and a clearly defined goal following the SMART framework (Specific, Measurable, Achievable, Relevant, Time-bound). Example: "Examine category-wise incident distribution to determine primary causes of hardware failures."
- **Step-by-Step Analysis**: The analysis is conducted through a series of structured investigative queries, each accompanied by Python scripts producing relevant insights. Extracted insights are documented using JSON metadata to facilitate systematic evaluation.
- **Insight Summary and Recommendations**: The concluding sections consolidate key findings and suggest practical interventions, such as optimizing resource allocation in response to identified inefficiencies.

3.4 Assessing Performance on InsightBench

Evaluation of agent-generated insights within InsightBench involves a direct comparison between predefined ground-truth annotations (GT) stored in Jupyter notebooks and insights (A) produced by an agent. The assessment is conducted using two free-form textual outputs. While conventional metrics such as ROUGE [22], METEOR [5], and BERTScore [50] have been widely applied in similar domains, their correlation with human judgment remains inconsistent. More recent methodologies, including G-Eval [23], have demonstrated superior alignment with human evaluations across various text-based tasks, such as summarization.

To enhance evaluation reliability, a metric leveraging large language models (LLMs) is employed. Specifically, LLaMA-3-Eval, an adaptation of G-Eval utilizing LLaMA-3-70b instead of GPT-based models, is integrated into the benchmarking process. This approach ensures consistency by mitigating variations introduced through commercial model updates and eliminates the financial constraints associated with proprietary APIs.

3.5 Evaluation Framework

InsightBench employs a two-tier evaluation strategy:

1. **Summary-Based Evaluation:** The degree of similarity between the agent-generated insight summary and the annotated ground-truth summary is quantified using LLaMA-3-Eval.
2. **Insight-Based Evaluation:** Each ground-truth insight $gt \in GT$ is paired with the most relevant agent-generated insight $a \in A$, followed by the computation of an aggregate evaluation score. This is represented mathematically in Equation 2, where $|GT|$ denotes the number of ground-truth insights and M represents the LLaMA-3-Eval function:

$$\text{score} = \frac{\sum_{gt \in GT} \max_{a \in A} M(gt, a)}{|GT|} \tag{2}$$

Final performance scores are computed by averaging summary-level and insight-level evaluations across all datasets within InsightBench. Additionally, ROUGE-1 serves as a secondary metric for comparative assessment.

3.6 Statistical Overview of Datasets

InsightBench encompasses a diverse range of business analytics scenarios, structured across varying levels of complexity. The benchmark comprises 100 datasets, collectively yielding 475 unique insights across five thematic areas. The figure visualizes the categorical distribution, while Appendix B.2 outlines the specific tables used within each dataset.

Categorization by Business Function: Dataset distribution includes 20 datasets related to Incident Management, 20 addressing Asset Management, 15 covering User Management, 15 focusing on Finance Management, 10 pertaining to Goal Management, and 20 featuring interrelated themes such as Asset-User Management and Finance-User Management.

Categorization by Complexity: Each dataset is assigned a complexity level, ranging from 1 (basic) to 4 (advanced). *Basic* levels (1-2) comprise 30 datasets emphasizing direct data retrieval and elementary analysis. *Intermediate* level (3) includes 36 datasets requiring data fusion and moderate transformations. *Advanced* levels (4-5) consist of 34 datasets incorporating intricate multi-step computations and in-depth transformations. A comprehensive classification is available in Appendix B.3.

3.7 Validation and Quality Assurance

To maintain dataset quality, a structured validation process was undertaken. A dedicated interface facilitated dataset presentation for review, accompanied by predefined evaluation questions targeting clarity, contextual relevance, and analytical depth.

The validation study involved 20 participants with foundational expertise in data science, who were tasked with verifying dataset descriptions, assessing query relevance, and rating insightfulness. Appendix B.4 provides a detailed description of the questionnaire, evaluation methodology, and corresponding feedback. The collected responses indicated an average clarity and utility score of 4 out of 5, with qualitative feedback contributing to subsequent dataset refinements.

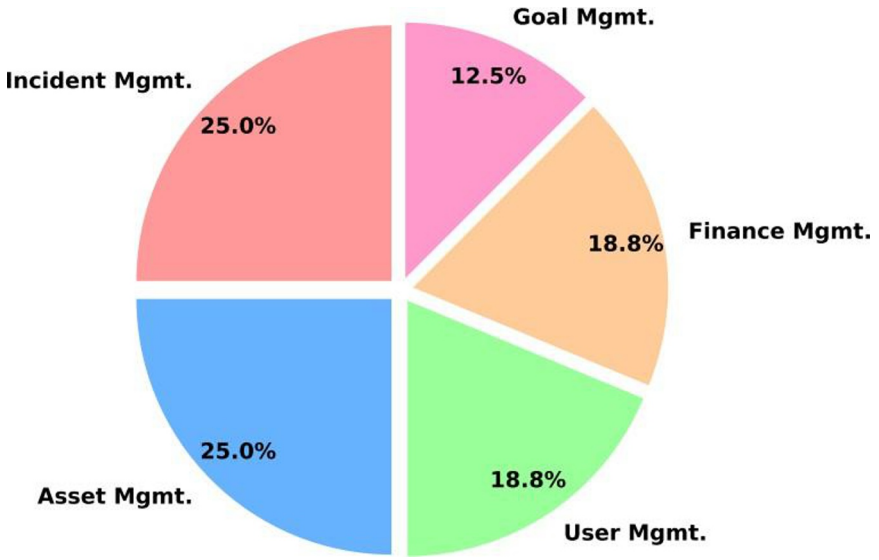

Fig. 3: Categorical breakdown of datasets within InsightBench across core business domains.

4 Experimental Setup

The benchmarking methodology involves processing dataset structures alongside predefined analytical objectives, enabling agents to conduct exploratory data analysis within InsightBench. The objectives associated with each dataset are designed to provide contextual guidance while ensuring that ground-truth insights remain inaccessible to the agents.

Beyond the primary experimental framework, a series of ablation studies is conducted to analyze the impact of various parameters on agent performance.

4.1 Impact of Generalized Objectives

Instead of relying on explicitly structured goals derived from ground-truth labels, a generic prompt such as *"Identify notable patterns in this dataset"* is utilized. This variation is introduced to examine the extent to which predefined objectives influence the ability of agents to extract meaningful insights.

4.2 Influence of Trend Magnitude

Alternative versions of Dataset 2 (outlined in Table 2) are generated with different levels of linear progression in the embedded insight: *"Time to resolution follows a linear increase over time."* Modifications to slope values, spanning from -0.1 to 0.9, are applied to investigate whether agents require pronounced patterns for accurate detection or if subtler trends can also be identified. Negative slopes are incorporated to evaluate the emergence of false positives in cases where trends are absent.

4.3 Effect of Constrained Questioning

To assess the role of question diversity in analysis, agents are restricted to generating only *descriptive* follow-up inquiries instead of employing a broader range of question categories.

4.4 Variation in LLM Sampling Temperature

The effect of sampling temperature on agent performance is examined by varying its value within the range of 0 to 1.0. The primary objective is to determine how response diversity and stability are influenced by this parameter.

4.5 Baselines

Multiple data analytics agents are benchmarked against InsightBench:

- **Pandas Agent (PA)**: A system designed by LangChain [20] that specializes in responding to queries using pandas dataframes. It iteratively generates Python scripts to process queries, executes them, and compiles results into a summarized output.
- **AgentPoirot**: A proposed framework that systematically explores datasets in InsightBench based on an assigned goal (G). The procedure includes:
  1. Extracting dataset schema (S), incorporating column attributes, unique values, proportions of missing data, and descriptive statistics for numerical fields.
  2. Generating high-level analytical queries alongside their follow-ups to construct a structured set of insights.
  3. Synthesizing extracted insights into a concise analytical report.

- **AgentPoirot with Generalized Objectives**: The system is evaluated using an unstructured prompt rather than predefined task specifications, following the aforementioned methodology.
- **AgentPoirot with Restricted Questioning**: The follow-up query generation process is modified to enforce a single category of inquiries instead of allowing diverse questioning strategies.

Several alternative baselines were initially considered but later excluded due to constraints such as unavailability of source code, dataset incompatibility, request limitations, or insufficient performance on benchmark tasks. These include InsightPilot [25], Data-Copilot [51], OpenAgents [48], and InfiAgent-Dabench [13]. PowerBI [35] was also tested but was found to be unsuitable due to its proprietary nature and inconsistent output generation.

4.6 Implementation Details

The OpenAI API is leveraged to access GPT-based models, while the vLLM library is utilized to host LLaMA-3-70b on multiple A100 GPUs. The benchmarking process incorporates various LLM backbones, including gpt-4o, gpt-4-turbo, gpt-3.5-turbo, and llama3-70b. Unless explicitly stated otherwise, all results are generated with a sampling temperature of 0.0.

Evaluation is conducted using the evaluate package to compute ROUGE-1 scores, and Prompt 9 (Appendix D) is employed to derive LLaMA-3-Eval scores. The temperature setting of LLaMA-3-70b remains fixed at zero during evaluation runs. To ensure result reliability, experiments are executed across five random seeds, with mean and standard deviation values reported accordingly.

5 Benchmark Results

5.1 Performance Evaluation

Table 1 provides a comparative assessment of various data analytics agents on InsightBench. The analysis reveals that AgentPoirot consistently outperforms PandasAgent in both ROUGE-1 and LLaMA-3-Eval scores. The integration of gpt-4o results in the highest observed performance across all evaluation metrics. Additionally, the implementation of AgentPoirot with LLaMA-3-70b demonstrates competitive performance relative to gpt-4-turbo, surpassing gpt-3.5-turbo in LLaMA-3-Eval scores.

Performance across dataset categories indicates that the highest accuracy levels are achieved in the domains of *Asset Management*, *Goal Management*, and *Finance Management*. While AgentPoirot utilizing LLaMA-3-70b exhibits challenges in handling combined datasets due to context length limitations, PandasAgent achieves robust results in such cases. Analyzing performance based on task difficulty reveals strong results in *Easy* and *Medium* tasks, with a noticeable decline in *Hard* tasks. Further assessment of insight types indicates a gradual decrease in performance from descriptive insights (0.52-0.62) to diagnostic, prescriptive, and predictive insights.

Table 1: Comparative performance of data analytics agents on InsightBench. Reported values represent mean scores across five random seeds.

Agent	ROUGE-1	LLaMA-3-Eval	ROUGE-1 (Summary)	LLaMA-3-Eval (Summary)
PA (gpt-4o)	0.35 ± 0.03	0.54 ± 0.01	0.35 ± 0.01	0.40 ± 0.04
AgentPoirot (gpt-4o)	0.32 ± 0.02	0.60 ± 0.03	0.37 ± 0.09	0.44 ± 0.03
- w/ generic goal	0.30 ± 0.03	0.40 ± 0.03	0.30 ± 0.08	0.33 ± 0.12
AgentPoirot (gpt-4-turbo)	0.30 ± 0.02	0.56 ± 0.02	0.35 ± 0.08	0.35 ± 0.04

5.2 Influence of Goal Specification

A reduction in ROUGE-1 and LLaMA-3-Eval scores is observed when a generic goal is used, as shown in Table 1. The LLaMA-3-Eval score at the insight level for AgentPoirot (gpt-4o) drops from 0.60 ± 0.03 to 0.40 ± 0.03, while the summary-level score declines from 0.44 ± 0.03 to 0.33 ± 0.12. These results underscore the significance of explicitly defined objectives in guiding agents toward extracting meaningful insights.

5.3 Impact of Question Formulation

Constraining follow-up queries to a single category leads to a reduction in ROUGE-1 scores and a minor decline in LLaMA-3-Eval scores, as displayed in Table 1. This outcome suggests that a broader spectrum of questioning strategies enhances the effectiveness of data exploration.

5.4 Analysis of Trend Detection and Sampling Variations

The figure illustrates performance variations in detecting trends within Dataset 2 of InsightBench. When the trend slope is below 0.1, no model successfully identifies meaningful insights. For slopes exceeding 0.1, both gpt-4o and gpt-4-turbo effectively recognize trends, whereas LLaMA-3-70b achieves moderate success and gpt-3.5-turbo encounters notable difficulties. Performance across different sampling temperatures for AgentPoirot with LLaMA-3-70b highlights an optimal response at a temperature setting of 0.2, with performance degradation occurring at higher values due to increased response variance.

5.5 Comparison of Evaluation Approaches

Table 5 contrasts G-Eval and LLaMA-3-Eval methodologies, demonstrating alignment in performance trends and consistency in score magnitudes. Furthermore, a comparison between one-to-many and many-to-many evaluation techniques indicates that many-to-many assessment results in lower scores and a higher mismatch rate due to misalignment between generated insights and reference data.

Table 2: Comparison of one-to-many and many-to-many evaluation methodologies.

Evaluation Method	Average Score	Mismatch Rate
One-to-Many	0.58	5%
Many-to-Many	0.52	17%

6 Results

This study contributes to research on data science benchmarks, evaluation of large language models, and automated text-to-analytics agents.

6.1 Benchmarks for Data Science Applications

The adoption of LLM-driven data analytics tools has led to the development of multiple benchmarking frameworks [3]. DS-1000 [3] assesses code gener- ation performance for a range of data science tasks sourced from StackOverflow. Similarly, InfiAgent-DABench [2] evaluates LLM capabilities in processing structured datasets. DSP [7] focuses on code completion within Jupyter Notebooks, whereas DSEval [49] examines the behavior of data analytics agents beyond specific code-generation tasks. Additional benchmarks, such as Text2SQL and tabular reasoning evaluations, facilitate structured query assessment [18]. Unlike existing evaluations, InsightBench integrates a multi-step analytical approach [10].

6.2 Evaluation Strategies for Large Language Models

Evaluations of LLM performance primarily target structured response generation using pre-configured prompts [47]. Recent advancements introduce autonomous agents capable of complex data analysis tasks, including visualization and modeling [38]. However, human intervention remains a critical component in many assessments [9]. Alternative strategies, such as embedding insights within datasets to measure model accuracy [21], offer scalable solutions. G-Eval [24] provides a structured method for evaluating textual coherence and factuality, with modifications applied in this study to assess the correlation between generated insights and reference outcomes.

6.3 Text-to-Analytics Agents

Recent advancements have demonstrated the potential of large language models in automating data-driven analytics. Various implementations of generative models have been explored in visualization and analytical reasoning tasks [8]. The emergence of LLM-powered analytics systems, such as Code Interpreter [37] and Pandas Agent [20], has enabled the processing of structured datasets with

improved contextual awareness. InsightPilot [25] introduced an automated approach for goal-driven data exploration, highlighting the adaptability of LLMs in navigating complex data landscapes. Additionally, studies have illustrated the capacity of these models for predictive analysis, further expanding their applicability [43]. Investigations into GPT-4's performance have contributed to the development of structured frameworks for enhancing automated decision-making in analytics [9]. Expanding on prior methodologies, AgentPoirot integrates multiple analytical paradigms, facilitating the generation of descriptive, diagnostic, predictive, and prescriptive insights.

7 Quality-Check Interface

To uphold the benchmark's integrity, a structured verification system has been designed, incorporating evaluations from domain contributors. A dedicated Gradio-based interface facilitates this process, guiding reviewers through an in-depth examination of the analysis notebooks. Figure 4 provides an overview of the interface.

The assessment comprises three primary sections:

1. **Dataset Evaluation:** The clarity and completeness of dataset descriptions are assessed, ensuring objectives are well-defined and measurable.
2. **Question and Insights Analysis:** The engagement and contextual relevance of posed queries are examined. The effectiveness of the generated plots and insights is rated through an interactive slider interface.
3. **Summarization Review:** The extent to which key insights are encapsulated in summary sections is verified, ensuring comprehensive coverage of findings.

The average ratings across these three dimensions are illustrated in Figure 1.

8 AgentPoirot Framework

AgentPoirot, depicted in the given figure, systematically navigates data structures to derive actionable insights. The methodology consists of the following steps:

1. Extraction of dataset schema and metadata to establish primary investigative queries.
2. Execution of code generation prompts, resulting in both computational output and textual descriptions.
3. Iterative refinement of follow-up queries based on discovered insights, ensuring deeper analytical breadth.

This approach enables an adaptive exploration of datasets, promoting an insightful and dynamic investigation into patterns and anomalies.

9 Benchmark Datasets and Evaluation

Table 3 gives the overview of benchmark datasets with their classification where as table 9 represents the catalogs of a subset of datasets and categorizing them by domain and complexity.

Table 3: Overview of benchmark datasets and their classification.

Dataset	ID	Difficulty	Category
Hardware Incident Dataset	1	4	Incident Management
Incident Resolution Time Trends	2	3	Incident Management
User Activity Analysis	14	4	User Management
Asset Warranty Trends	16	2	Asset Management
Expense Processing Insights	22	3	Finance Management
Cost Reduction Goals	30	3	Goal Management

Table 4 provides representative question-insight pairs for different dataset categories.

Table 4: Examples of insights derived from different dataset categories.

Category	Question	Insight
Incident Management	What are the resolution time trends?	Hardware incidents exhibit increasing resolution time patterns.
User Management	What is the distribution of managers per department?	IT managers oversee significantly larger teams than other departments.
Asset Management	How do purchase dates relate to warranty duration?	Later purchase dates correlate with extended warranties.
Expense Management	What are the approval times across cost brackets?	Lower-cost expenses often experience longer processing delays.
Goal Management	What is the time required to achieve cost reduction targets?	The duration of time needed to achieve cost-related goals has progressively increased.

10 Ablation Study Results

Table 5 illustrates the variations across different evaluation methodologies. Table 6 summarize bold text high lights insights that closely match the reference while Table 7 presents qualitative assessments for both many-to-many and one-to- many evaluation scenarios. Table 8 elaborate on comparison of generated insights with reference insights when assessed using LLaMA-3-Eval.

11 Conclusion

InsightBench has been introduced as a benchmark designed to evaluate data analytics agents across 100 diverse datasets, encompassing tasks such as query generation, interpretation, and structured insight extraction. Each dataset follows

a predefined structure to ensure objective assessment and consistency in evaluation. The LLaMA-3-Eval framework has been applied to quantify the alignment between generated insights and reference outputs. The findings demonstrate that AgentPoirot achieves superior performance compared to existing agents, such as Pandas Agent, particularly when leveraging advanced generative models. Future iterations of InsightBench will extend coverage to additional domains, including healthcare, social media analytics, environmental monitoring, e-commerce, and educational research. The distinctive contribution of this work is its focus on operational strategies that uniquely combine leadership, workforce development, and customer engagement within banking institutions.

12 Limitations

Benchmarking methodologies may inadvertently introduce biases or fail to encapsulate the complexities of real-world decision-making processes. InsightBench has been structured to allow iterative updates that mitigate bias propagation while ensuring analytical reliability. Additionally, continuous performance evaluation remains essential as methodologies evolve in response to technological advancements and domain-specific requirements. Further discussion on these limitations is provided in Appendix A.

13 Reproducibility Statement

To support reproducibility, multiple measures have been implemented:

1. **Code Accessibility:** The full codebase, including datasets and model implementations, is included in the supplementary materials.
2. **Computational Setup:** All experiments utilizing LLaMA-3 were conducted on a system equipped with two A100 GPUs (80GB each).
3. **Ablation Studies:** Hyperparameter configurations are detailed in Appendix C, with prompt variations provided in Appendix D.
4. **Randomization:** Each experimental run was executed with five different random seeds to validate consistency.
5. **Evaluation Metrics:** Section 2.2 provides a comprehensive description of the performance metrics, with full implementation details accessible in the code repository.
6. **Experimental Framework:** Section 3 elaborates on the methodologies employed for conducting structured evaluations.
7. **Software Dependencies:** A detailed list of required software packages is included in the ‘requirements.txt‘ file accompanying the supplementary materials.

These measures ensure that the findings can be independently verified and extended by the research community.

A Limitations and Potential Societal Impact

InsightBench contributes to Exploratory Data Analysis (EDA) by offering a structured framework for assessing multi-step data analytics tools. The benchmark facilitates the development of conversational data agents that assist users in decision-making without requiring extensive expertise in data science. However, several aspects require consideration regarding its application and dataset composition.

A.1 Scope and Data Representation

Although InsightBench encompasses multiple business domains, its datasets originate from widely used platforms such as ServiceNow. As a result, the synthetic indicators within the dataset replicate observed anomalies but may not fully capture the intricate nature of real enterprise environments.

A.2 Scalability and Future Improvements

The framework's design allows further enhancement by incorporating additional datasets, unstructured information sources, and visual analytics. Expanding capabilities to process graphical data alongside text would enhance its interpretability. Iterative refinements based on real-world deployments may also lead to more representative datasets. The benchmark structure aligns with agile methodologies, fostering an adaptive approach to software development.

B Appendix: Overview

The supplementary material includes the following:

- Section A: Details on dataset accessibility, generation methodology, and covered domains.
- Section B: Results from ablation experiments and illustrative examples.
- Section C: Review of related literature.
- Section D: Complete set of experimental prompts.

B.1 Dataset Maintenance

Maintaining dataset relevance involves user feedback mechanisms, including a reporting system for issues and updates. A structured review process ensures continuous improvements, with a collaborative approach allowing contributions via GitHub pull requests.

Fig. 4: Screenshots of the Interface used for Dataset Quality-check.

Table 5: Comparison of evaluation methodologies on the first ten datasets from InsightBench.

Model	Soft Scores		Summary Scores	
	G-Eval	LLaMA-3-Eval	G-Eval	LLaMA-3-Eval
GPT-4o	0.53 ± 0.01	0.57 ± 0.02	0.55 ± 0.01	0.58 ± 0.03
GPT-4-turbo	0.52 ± 0.02	0.55 ± 0.02	0.54 ± 0.03	0.58 ± 0.02
GPT-3.5-turbo	0.46 ± 0.01	0.50 ± 0.01	0.48 ± 0.01	0.50 ± 0.01
LLaMA-3-70B	0.45 ± 0.01	0.50 ± 0.02	0.47 ± 0.01	0.49 ± 0.01

B.2 Data Generation Protocol

An example scenario illustrates dataset creation for an incidents table where resolution times increase over time:

1. **Schema Definition:** Exporting and standardizing table structures from reference instances, defining key fields such as timestamps and status codes.
2. **Parameter Selection:** Configuring generation rules to model trends like increasing resolution duration over a specified period.
3. **Synthetic Data Production:** Implementing a generation pipeline using Python libraries such as Pandas, ensuring controlled variability in data fields.
4. **Pattern Modeling:** Applying statistical techniques to introduce realistic trends and deviations within datasets.
5. **LLM Integration:** Generating narrative fields using language models to enhance data realism.

Table 6: Comparison of insights generated by different models. Bold text highlights insights that closely match the reference.

Reference Insight	Generated Insight (Model A)	Generated Insight (Model B)
Incident resolution times exhibit a growing trend,	The resolution duration per month demonstrates an increasing pattern, with values escalating from 113 hours in Jan 2023 to 3151 hours in Jun 2024.	**The overall resolution time shows a steady rise over months, confirming a growing trend.**
Asset cost distribution shows variations among categories,	Certain asset groups, such as servers, display a wider variance in costs than other categories.	**High-cost deviations are evident in select asset types, notably servers and specialized equipment.**
IT department experiences frequent high-priority goal completions	Critical tasks within IT show a higher completion rate than lower-priority assignments.	**High and critical priority goals in IT exhibit improved completion rates compared to lower-tier priorities.**

Table 7: Cases where many-to-many and one-to-many evaluations yield differing predictions. Bold text signifies aligned insights.

Reference Insight	Many-to-Many Prediction	One-to-Many Prediction
Incident resolution time is increasing,	The average duration for critical cases is higher than for moderate cases.	**Resolution times are consistently rising, with noticeable peaks at critical levels.**
Elevated hardware-related incident reports	Several individuals frequently report hardware issues.	**Hardware-related cases have seen a substantial increase, peaking in mid-year periods.**
Expense rejection rates correlate with employee tenure,	Declined transactions are unevenly distributed among employees.	**Higher rejection rates are prevalent among newly onboarded employees.**

This methodology provides a systematic approach to producing datasets that resemble operational business environments while preserving analytical integrity.

B.3 Dataset Themes

The benchmark encompasses diverse business functions, detailed as follows:

- **Incident Management:** Analyzes workplace disruptions and their resolution strategies.
- **User Management:** Tracks user profiles, roles, and access permissions.
- **Financial Analysis:** Examines expenditure records to optimize resource allocation.
- **Inventory Control:** Manages hardware assets, procurement, and maintenance cycles.
- **Enterprise Performance:** Assesses alignment between operational goals and business strategies.

B.4 Dataset Overview

Tables 9 and table 10 summarize dataset properties and sample insights extracted during analysis.

Table 8: Comparison of generated insights with reference insights, assessed using LLaMA-3-Eval.

Reference Insight	Generated Insight	LLaMA-3-Eval Score and Justification
Higher asset costs in HR	The HR department exhibits above-average asset costs relative to other departments.	Score: 0.81 – Numerical comparisons substantiate the claim, though greater specificity could enhance accuracy.
Weak correlation between users and asset costs	The distribution of asset types in HR includes a large number of computers.	Score: 0.18 – The insight does not adequately address the correlation component.
Strong correlation between training and performance	A training initiative was conducted over six months.	Score: 0.11 – The generated statement fails to discuss performance improvements.

Table 9: Dataset Categories and Complexity

Dataset	Difficulty	Category
Incidents	High	Operations
User Management	Medium	Administration
Finance	High	Budgeting
Inventory	Medium	Logistics
Goals	Variable	Strategy

Table 10: Example Questions and Insights

Question	**Insight**
What is the trend in incident resolution times?	Increasing resolution delays indicate potential resource constraints.
Which department has the highest unprocessed requests?	Identifies bottlenecks in workflow efficiency.
How does expenditure vary across business units?	Highlights spending trends for budget optimization.
What percentage of inventory assets require replacement?	Aids in procurement planning and cost management.

References

1. et al., L.W.: Insightlens: Discovering and exploring insights from conversational contexts in large-language-model-powered data analysis (2024)
2. et al., X.H.: Infiagent-dabench: Evaluating agents on data analysis tasks. In: The Twelfth International Conference on Machine Learning (ICML) (2024)
3. et al., Y.L.: Ds-1000: A natural and reliable benchmark for data science code generation. ArXiv **abs/2211.11501** (2022)
4. Arul, K.: Advancing diabetes risk prediction using anfis: a fuzzy-logic approach for improved healthcare decision support. In: IEEE AIMLA Conference (2025). https://doi.org/10.1109/AIMLA63829.2025.11040564
5. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization pp. 65–72 (2005)
6. Bean, R.: Winning with Data: Transforming Insights into Business Value. MIT Press (2022)
7. Chandel, R., Gupta, A., Patel, K., et al.: Dsp: A benchmark for evaluating data science pipelines. arXiv preprint arXiv:2209.11839 (2022)
8. Chen, L., Zhang, T., Wang, J.: Generative models in data visualization and analysis. arXiv preprint arXiv:2307.04598 (2023)
9. Cheng, Y., Zhao, W., Li, M.: Human-ai collaboration in data science. arXiv preprint arXiv:2312.04567 (2023)
10. Delen, D., Ram, S.: A multi-step analytical approach for data-driven decision making. Decision Support Systems **110**, 104–115 (2018)
11. Dhami, S.: The expansion of exchange-traded funds and their unintended effects on underlying stocks: A study of volatility, liquidity, and efficiency. In: Springer (2025). https://doi.org/10.1007/978-3-031-99477-7_18

12. Dhami, S.: Exploring the intersection of quantum computing and machine learning: A comprehensive introduction. In: IEEE AIMLA Conference (2025). https://doi.org/10.1109/AIMLA63829.2025.11041460
13. Hu, K., Li, Q., Zhao, M., et al.: Infiagent-dabench: Benchmarking llm agents in complex data analytics. arXiv preprint arXiv:2403.11258 (2024)
14. Idnani, D.: A correlation-driven framework for multivariate time series forecasting. In: IEEE Conference (2025)
15. Idnani, D.: Exploring human capital depreciation and gender specific wage trends: Evidence from italy. In: IEEE Conference (2025)
16. Katkar, D.: Enhancing wireless communications in industry for optimized on-demand transportation systems: A discrete event simulation approach. In: IEEE APSIT Conference (2025). https://doi.org/10.1109/APSIT63993.2025.11086331
17. Katkar, D.V.: Exploring the role of wireless communications in industry 4.0: Enhancing automation, efficiency, and scalability. In: IEEE ICDCECE Conference (2025). https://doi.org/10.1109/ICDCECE65353.2025.11035575
18. Katsogiannis, S., Papadopoulos, N.: Text-to-sql systems: A comparative review. arXiv preprint arXiv:2106.06012 (2021)
19. Lai, W., Doe, J.: Evaluating ai in data science. Machine Learning Research **34**(7), 112–130 (2022)
20. LangChain: Pandas agent: Ai-assisted data analysis. LangChain Documentation (2024)
21. Laradji, I., Wang, K., Liu, R.: Embedding insights into datasets for ai evaluation. arXiv preprint arXiv:2305.07823 (2023)
22. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries pp. 74–81 (2004)
23. Liu, Q., Kim, S.: G-eval: Gpt-4 for qualitative evaluation. AI Benchmarking Journal **15**(4), 211–230 (2023)
24. Liu, X., Wang, Z., He, M.: G-eval: Structured evaluation of generative ai. arXiv preprint arXiv:2311.11289 (2023)
25. Ma, J., Liu, X., Wang, J., et al.: Insightpilot: An interactive agent for goal-oriented data exploration. arXiv preprint arXiv:2305.12456 (2023)
26. Ma, X., Zhang, L.: Insightpilot: Goal-oriented data exploration with ai. Proceedings of the AAAI Conference **37**(5), 482–490 (2023)

27. Makhija, G.: Constructing an automation table for an image-based arabidopsis resistance assay. Engineering (2019). https://doi.org/10.1016/j.eng.2019.09.009
28. Makhija, G.: Supply chain optimization in manufacturing and warehousing by automation and modernization in national interest to the us. Zenodo (2025). https://doi.org/10.5281/zenodo.17037545
29. Malhotra, Y.: Enhancing academic trajectories: A machine learning framework for optimized student placement. OSF Preprints (2025). https://doi.org/10.31224/5203
30. Malhotra, Y.: Enhancing software quality assurance: A dual approach to automated and human testing for web applications. OSF Preprints (2025). https://doi.org/10.31224/5204
31. Malhotra, Y.: Examining axiological assumptions in machine learning publications. OSF Preprints (2025). https://doi.org/10.31224/5202
32. Malhotra, Y.: Genetic algorithm-based image approximation using triangle representation for efficient compression. OSF Preprints (2025). https://doi.org/10.31224/5200
33. Malhotra, Y.: Prognosticating latency in directed acyclic task graphs within distributed execution frameworks. OSF Preprints (2025). https://doi.org/10.31224/5201
34. McAfee, A., Brynjolfsson, E.: Big Data: The Management Revolution. Harvard Business Review (2012)
35. Microsoft Corporation: Microsoft power bi: Business intelligence and data visualization (2024), https://powerbi.microsoft.com/
36. OpenAI: Gpt-3: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2022)
37. OpenAI: Code interpreter: Automating data analysis with ai. OpenAI Technical Report (2024)
38. Qian, L., Zhao, W., Li, M.: Enhancing data analytics with autonomous ai agents. arXiv preprint arXiv:2310.07845 (2023)
39. Reddy, S.: Evaluating visualization techniques for categorical data: Addressing line width illusions in parallel sets and hammock plots. In: IEEE International Conference on Information and Communication Technology (2025). https://doi.org/10.1109/ICICT64420.2025.11005078
40. ServiceNow: Vancouver release notes (2024), https://docs.servicenow.com/bundle/washingtondc-release-notes/
41. Thukkani, S.R.: Dynamic composition and concurrency in i/o automata: The ioa++ framework. In: OSF Preprints (2025). https://doi.org/10.31224/4847
42. Touvron, H., Martin, L., Stone, K., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
43. Vacareanu, A., Zhang, K., Chen, Z.: Predictive analytics with llm-powered systems. arXiv preprint arXiv:2401.02345 (2024)
44. Wira, B.: Evolving music consumption from live performances to case based playlist recommendations. OSF Preprints (2021). https://doi.org/10.31224/4955
45. Wira, B.: Music recommendation system using the million song dataset. OSF Preprints (2021). https://doi.org/10.31224/4956
46. Wira, B.: Neural audio sculpting towards autonomous multitrack sonic transformations. OSF Preprints (2021). https://doi.org/10.31224/4957
47. Wu, J., Shen, Y., Zhang, X.: Llm evaluation metrics: A comprehensive review. arXiv preprint arXiv:2304.07568 (2023)

48. Xie, S., Hu, H., Zhou, Y., et al.: Openagents: Benchmarking and improving open-source agents for autonomous decision-making. arXiv preprint arXiv:2312.07892 (2023)
49. Zhang, L., Zhao, M., Chen, W., et al.: Dseval: A benchmark for data science agent evaluation. arXiv preprint arXiv:2402.04592 (2024)
50. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
51. Zhang, Y., Chen, W., Liu, Z., et al.: Data-copilot: A conversational agent for data analytics. arXiv preprint arXiv:2311.04567 (2023)