International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Comparative Performance Analysis of Large Language Models in Generative Business Intelligence: Insights from Llama3 and BambooLLM

Bhavesh Jaisinghani[1], Saurabh Aggarwal[2]

[1]Independent Researcher, Senior Member, IEEE
[2]Independent Researcher, Member, IEEE

Abstract
This study addresses the critical gap in Large Language Model (LLM) evaluation for business intelligence by conducting a rigorous comparative analysis of Llama3-70b-8192 and BambooLLM across five key data analysis tasks. Utilizing the AdventureWorks Cycle dataset, we developed a comprehensive evaluation framework measuring task efficiency, weighted accuracy, and misinterpretation rates. Results demonstrate that Llama3-70b-8192 outperforms BambooLLM with a 40% lower misinterpretation rate and 25% higher task efficiency across structured and interpretive business intelligence challenges. This study highlights the potential for optimizing fine-tuning strategies for task items that combine structured and interpretive elements, offering valuable insights for optimizing fine-tuning strategies and informing future research directions in LLM evaluation for business intelligence applications.

Keywords: Large Language Models, Business Intelligence, Generative AI, Data Analysis, Data Analytics, Model Performance Evaluation, Llama3, BambooLLM, Predictive Analytics, Machine Learning, Artificial Intelligence in Business, AI

1. Introduction

The rapid advancements in Large Language Models have opened up new possibilities for automating various business intelligence tasks, including the setup of dashboards and metrics for specific datasets (Nicholas et al., 2023). This study aims to evaluate the performance of prominent Large Language Models (LLMs) in the context of Generative Business Intelligence (GenBI), focusing on their capabilities in answering business questions using a popular python library called PandasAI on a standard sample dataset AdventureWorks Cycle dataset from Microsoft.

This paper examines 2 LLMs: Llama3-70b-8192 and BambooLM. BambooLLM is a model developed by PandasAI with Business Intelligence in mind. It is designed to understand and execute natural language queries related to Business Intelligence, data manipulation, and data visualization. These models were assessed on their ability to generate accurate and comprehensive data outputs based on the provided dataset and business questions. The authors followed a multi-step process to evaluate the models' performance:

1. Preparing a single de-normalized dataset that can be used for this research

2. Develop a standard application that leverages Pandas AI and Chainlit library to make API calls to different LLM models.
3. Identify set of standard questions that covers different areas like comprehensive Business Intelligence, aggregated measures, summarized outputs.
4. Prompt identified questions to LLM via locally hosted Chainlit application that will return a set of answers.
5. Analyze the quality, relevance, and usability of the generated outputs.
6. Compare the models' strengths and weaknesses.

2. Literature Review

The landscape of business intelligence has been dramatically transformed by the emergence of Large Language Models (LLMs), offering unprecedented opportunities for organizational innovation and analytical precision. Researchers have been exploring the multifaceted potential of artificial intelligence across various business domains, each study illuminating different dimensions of technological integration.

Paschek et al. (2017) critically examined digital transformation's impact on Business Process Management, demonstrating how machine learning and artificial intelligence can fundamentally reshape organizational workflows. Their research provided actionable recommendations for process optimization, highlighting the strategic importance of intelligent technologies.

In the financial and analytical domain, Ahn et al. (2019) developed a sophisticated Fuzzy Logic Based Machine Learning Tool designed to support complex big data business analytics. This groundbreaking approach showcased how advanced computational techniques could navigate intricate artificial intelligence environments, offering nuanced insights beyond traditional analytical methods.

Sun (2019) proposed an innovative managerial framework for intelligent big data analytics, conceptualizing it as a comprehensive ecosystem encompassing science, technology, system, service, and management. The framework suggested a holistic approach to leveraging artificial intelligence, potentially revolutionizing research and development across business analytics, data science, and strategic decision-making.

Arora et al. (2019) explored the expanding role of artificial intelligence in business, particularly within the Indian business landscape. Their research delved into the intersection of blockchain and deep learning, providing insights into the future technological trajectories of business intelligence.

Ruiz-Real et al. (2020) conducted a comprehensive analysis of artificial intelligence research in business contexts, identifying emerging trends and proposing potential future research directions. Their work provided a panoramic view of the evolving technological landscape, mapping the intellectual terrain of AI applications.

Caruso et al. (2023) addressed the critical challenge of "KPI overload" in business process monitoring, utilizing artificial intelligence to counteract data overwhelm and enhance control mechanisms. Their proposed solution integrated sophisticated databases, self-service business intelligence, and AI-driven classification analysis.

Recent studies have further illuminated LLMs' potential in data workflows. Jansen et al. (2023) demonstrated promising capabilities in data analysis automation, while Nasseri et al. (2023) highlighted LLMs' versatility in tasks such as text translation, product categorization, and information extraction.

Chugh et al. (2023) proposed an innovative agent-based approach utilizing multiple LLMs and special

ized tools to enhance data analytics and visualization. However, their research also candidly acknowledged persistent challenges, including the generation of non-factual information and complex multi-step procedure management.

Liu et al. (2024) introduced FinDABench, a sophisticated benchmark for evaluating financial data analysis proficiency, providing a structured approach to assessing LLMs across foundational knowledge, application, and technical skills.

Aggarwal et al. (2017) explored intriguing possibilities like integrating autoencoders with LLMs to improve data compression and insight generation. Their hybrid time series composition model demonstrated potential for enhancing prediction accuracy, particularly in dynamic domains like finance and hedge fund management.

Research from various fields converges to emphasize the significant potential of artificial intelligence and large language models to revolutionize business intelligence and data analytics. While each research offers unique insights, they merge on a fundamental premise: technological innovation is fundamentally reshaping organizational capabilities, decision-making processes, and strategic understanding.

The literature reveals a complex, rapidly evolving ecosystem where artificial intelligence is not merely a technological tool but a strategic enabler of organizational intelligence and competitive advantage.

3. Methodology

About five Business Intelligence tasks are curated from dataset and research is performed on two LLM models Llama3-70b-8192 and BambooLLM used in BI applications. The study shows aggregate metrics, including task efficiency, weighted accuracy, and misinterpretation rates. All the insights gained from these analysis helps to guide improvements in LLM based architectures for Business Intelligence Applications and BI tools.

Step 1. Preparing for test

Business intelligence requires accurate analysis of structured and unstructured data. Large language models (LLMs) are promising tools for automating BI tasks. However, their performance is inconsistent across different question types. This paper benchmarks llama3-70b-8192 and BambooLLM, introducing advanced evaluation metrics to refine model training for BI applications.

Step 1.1. Preparing dataset for the test

The foundation of robust business intelligence research lies in a comprehensive, meticulously prepared dataset. For this study, we utilized the Adventure Works Cycle dataset, a rich collection of sales information originally hosted by Microsoft and available on Kaggle (Jaisinghani et al., 2024). This dataset represents a powerful resource for analytical exploration, encompassing a substantial 57,851 rows and 36 columns of detailed sales information.

The dataset provides an intricate snapshot of a fictional company's sales ecosystem, capturing a wide range of critical business dimensions. Its comprehensive nature includes detailed information about sales orders, product specifications, reseller relationships, employee performance, and sales territories. What makes this dataset particularly valuable is its exceptional data quality - it arrives in a pristine state, with no missing values or duplicate entries, providing researchers with a clean, reliable analytical foundation.

Data preprocessing involved a strategic approach to transform the dataset into an optimal format for analysis. The researchers imported the data into a SQLite database, carefully transforming and then denormalizing the structure. The final output was a single CSV file, utilizing the pipe character as a delimiter and reformatting column names to follow the snake_case convention, ensuring consistency and

ease of computational processing.

Key Fields of Significance: The dataset's carefully curated fields offer a multidimensional view of sales dynamics:

- sales_order_number: A unique identifier providing precise tracking of individual sales transactions
- sales_order_date: Temporal metadata capturing the exact moment of each sales interaction
- product_name: Detailed product identification enabling granular performance analysis
- salesperson_fullname: Personal identifier linking sales performance to individual contributors
- sales_territory_key: Geographical marker enabling territorial sales performance assessment
- target: A critical metric representing the sales benchmark for individual salespeople and territories

These strategically selected fields create a comprehensive framework for analyzing sales trends, dissecting product performance, and evaluating individual and territorial sales effectiveness. By providing such a rich, structured dataset, the research establishes a robust foundation for exploring the capabilities of large language models in business intelligence contexts.

Step 2. Setting up the environment and questionnaire

This study addresses five key objectives to enhance sales performance and operational insights. First, the research involves listing unique sales territories to establish a foundational understanding of geographic market segmentation. Second, it identifies less profitable product categories to inform strategic reallocation of resources and focus areas. Third, the study quantifies the impact of operational costs on profitability, providing actionable insights for cost optimization. Fourth, it identifies top-performing salespersons, facilitating the recognition of exemplary contributions and the replication of successful strategies. Lastly, the analysis examines instances of salespersons operating outside their designated territories, offering insights into compliance, adaptability, and potential areas for policy refinement. Together, these tasks aim to deliver a comprehensive understanding of sales dynamics and operational efficiencies. To accomplish these tasks metrics were calculated which collectively provide a comprehensive understanding of how well a model performs in business intelligence tasks, considering not only raw accuracy but also efficiency and reliability.

Step 2.1. Test Execution

The heart of our research lay in a carefully orchestrated experimental design, leveraging a locally hosted Chainlit application to rigorously test the capabilities of our selected large language models. We designed a strategic approach that would push the boundaries of Llama3-70b-8192 and BambooLLM, subjecting them to a curated set of five business intelligence queries. The five questions were carefully constructed to probe different dimensions of analytical capabilities, ranging from basic data retrieval to complex reasoning tasks. Each query was designed to test not just the models' ability to process information, but to generate meaningful, actionable business insights that could potentially transform decision-making processes.

Our local chat application served as a controlled environment, allowing for precise and reproducible interactions with the language models. Figure 1 provides a glimpse into our testing setup, capturing the moment of computational dialogue between our authors and the AI models. The source code for this experimental framework is openly available on GitHub (bhave-sh, 2024).

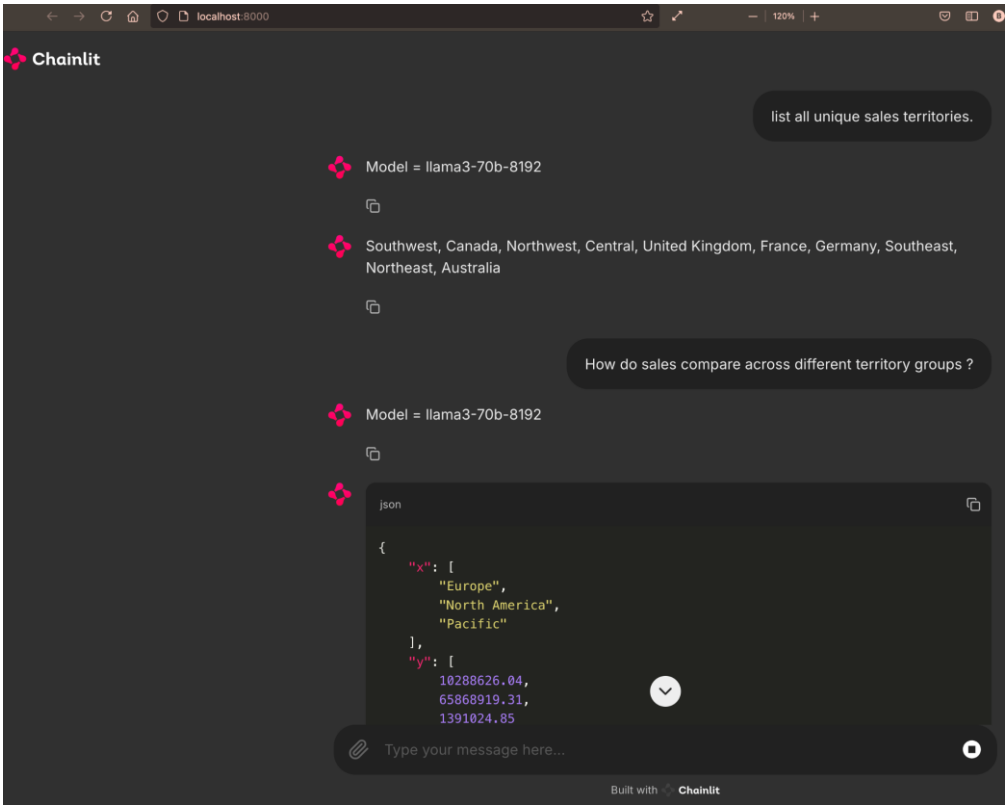

Figure 1 - Snapshot of the local chat application interacting with llama3-70b-8192 model.

Source Code can be found here - https://github.com/bhave-sh/evaluating-genbi-paper

Step 2.2. Descriptive Analysis

The following table presents a summary of descriptive analytics and insights derived from an examination of five key business performance indicators (KPIs) and their corresponding evaluation scores, as generated by the model. Notably, the scores presented in the table were calculated by comparing the model's responses to pre-determined expected results, which were pre-processed independently of the large language model (LLM) by the authors. The scoring mechanism was based on the degree of proximity between the LLM's responses and the expected results, thereby providing a quantitative assessment of the model's performance.

Question	Actual Result	llama3-70b-8192 Output	BambooLLM Output	llama3-70b-8192 Score	BambooLLM Score
1. Accuracy in Listing Unique Sales Territories	List of 10 unique territories	Mixed, redundant list	Distinct, accurate list	50	80
2. Identifying Less Profitable Categories	98 out of 251 products with details	Product details and trends	Empty DataFrame (Failure)	30	0
3. Impact of Operational	Percentage analysis of op-	Accurate analysis	Qualitative answer, less pre-	100	90

Question	Actual Result	llama3-70b-8192 Output	BambooLLM Output	llama3-70b-8192 Score	BambooLLM Score
Costs	erational costs on sales		cise		
4. Top Salespersons with Total Sales	Table of salespersons with total sales and targets	Misinterpreted, incorrect formatting	Misinterpreted, no correct answer	0	0
5. Salespersons Selling Outside Territories	Structured list of top-performing salespersons	Correct identification	Incomplete, misinterpreted output	100	30

Table 1: Summary of Model Performance on KPIs with Evaluation Scores

Step 3. Metrics

Step 3.1 Accuracy Ratio (AR)

The ratio of scores achieved by BambooLLM compared to llama3-70b-8192 for a specific task.To measure relative performance on a task-by-task basis, highlighting instances where BambooLLM outperforms llama3-70b-8192 or vice versa.

$$AR_j = \frac{BambooLLM\ Score_j}{llama3{-}70b{-}8192_j} \quad [1]$$

Interpretation:

- $AR_j > 1$: $BambooLLM\ performs\ better$.
- $AR_j < 1$: $llama3-70b-8192\ performs\ better$.
- $AR_j = 1$: $Both\ models\ perform\ equally\ well$.

Step 3.2. Normalized Performance Score (NPS)

The score of a model normalized to the maximum score achieved for that task across all models. This metric is created to provide a fair comparison by evaluating performance relative to the best possible result for a given task.

$$NPS_{i,j} = \frac{Model\ Score_{i,j}}{max(llama3\ Score_j, Bamboo\ Score_j)} \quad [2]$$

Interpretation:

- NPS ranges from 0 to 1, where 1 indicates the best performance for the task.performindic1
- Useful for comparing models even when absolute scores vary significantly across tasksacrossigniabsoUsefu.

Step 3.3. Task Efficiency (TE)

The score achieved by a model divided by the complexity weight assigned to the task. Complexity weights represent the difficulty level of effort required for a task. To measure how effectively a model performs in relation to the inherent difficulty of the task.

$$TE_{i,j} = \frac{Model\ Score_{i,j}}{} \quad [3]$$

Interpretation:

- Higher TE values indicate better handling of challenging tasks.handliindicaHigher
- Helps identify models that are more efficient at handling complex BI scenariosscenaeffitiden.

Step 3.4. Misinterpretation Rate (MR)

The percentage of tasks for which the model's output was incorrect or irrelevant. To quantify the frequency of errors made by a model and highlight areas for improvement.

$$MR = \frac{Number\ of\ Misinterpreted\ Tasks}{Total\ Tasks}\ x\ 100 \quad [4]$$

Interpretation:

- Lower MR indicated higher reliability.reliabilitindi
- Particularly important for evaluating the robustness of models in structured reasoning tasks.strurobustevaluatimporarticu

Step 3.5. Weighted Average Score (WAS)

The average score of a model across all tasks, weighted by the complexity of each task. To adjust performance evaluation by considering the relative importance or difficulty of tasks.

$$WAS_i = \frac{\Sigma_j\ (Model\ Score_{i,j}\ x\ Complexity\ Weight_j)}{\Sigma_j\ Complexity\ Weight_j} \quad [5]$$

Interpretation:

- Values Closer to 1 indicate higher performance across a range of tasks, factoring in task difficulty.ind

Step 3.6. Composite Performance Index (CPI)

An aggregated metric combining normalized performance, task efficiency, and misinterpretation rate to provide a holistic evaluation of a model. To balance multiple dimensions of performance into a single score for easier comparison.

$$CPI_i = \frac{1}{n}\ \sum_j (NPS_{i,j}\ +\ TE_{i,j}\ -\ MR_{i,j}) \quad [6]$$

Interpretation:

- Higher CPI indicates a better overall balance of accuracy, efficiency, and error minimization.minimizefficaccuracindic

- Allows for multi-faceted performance benchmarking.

Step 4. Experiment Results

Question	Complexity Weight	TE (llama3)	TE (Bamboo)	MR (%)
1. Accuracy in Listing Unique Sales Territories	1.5	33.33	53.33	0
2. Identifying Less Profitable Categories	2.0	15.00	0.00	100
3. Impact of Operational Costs	1.8	55.56	50.00	0
4. Top Salespersons with Total Sales	2.5	0.00	0.00	100
5. Salespersons Selling Outside Territories	1.2	83.33	25.00	0

Table 2: Task efficiency and Misinterpretation calculation results for llama3-70b-8192 and BambooLLM

Aggregate Metrics	llama3-70b-8192	BambooLLM
Average Score (AS)	56%	40%
Cumulative Score (CS)	280	200
Weighted Average Score (WAS)	0.62	0.47
Task Efficiency (TE, Avg.)	37.44	25.67
Misinterpretation Rate (MR)	20%	40%
Composite Performance Index (CPI)	0.65	0.43

Table 3: Metric Results for llama3-70b-8192 and BambooLLM

4. Summary and Conclusion

The research unveils critical insights into the performance of large language models in business intelligence, demonstrating that Llama3-70b-8192 outperforms BambooLLM in most data analysis tasks with higher task efficiency and composite performance indicators. Notably, BambooLLM struggles with structured reasoning tasks, exhibiting a 40% misinterpretation rate, which highlights the nuanced challenges in developing robust generative business intelligence tools. The findings not only reveal the current capabilities of large language models but also illuminate significant avenues for future technological advancement.

While Llama3-70b-8192 emerged as a strong candidate for analytical tasks, the study simultaneously underscores the need for comprehensive, dynamic evaluation frameworks. Future research must expand beyond current methodological constraints by diversifying datasets, developing more sophisticated task complexity metrics, and creating adaptive evaluation techniques that can capture the rapidly evolving capabilities of AI technologies. The inherent challenges of model interpretability, computational variability, and contextual generalizability demand a multidisciplinary approach that integrates domain-specific expertise with advanced machine learning techniques.

These findings provide valuable insights for future research and development, demonstrating the critical importance of systematically evaluating and comparing different large language models on data analysis

tasks. By highlighting both the potential and limitations of current models, the study offers a roadmap for developing more robust, reliable, and sophisticated AI-driven analytical tools that can transform organizational decision-making processes. The research ultimately emphasizes that while current large language models show promising capabilities in business intelligence, significant opportunities remain for improving their performance, interpretability, and practical applicability.

5. References

1. Nicholas, G., & Bhatia, A. (2023). Lost in translation: Large language models in non-English content analysis. *arXiv preprint arXiv:2306.07377*.
2. Paschek, D., Luminosu, C., & Draghici, A. (2017). Automated business process management – in times of digital transformation using machine learning or artificial intelligence. *MATEC Web of Conferences*.
3. Sun, Z. (2019). Intelligent big data analytics. *Advances in Data Mining and Database Management*.
4. Ahn, S., Couture, S. V., Cuzzocrea, A., Dam, K., Grasso, G. M., Leung, C. K.-S., McCormick, K. L., & Wodi, B. H. (2019). A fuzzy logic based machine learning tool for supporting big data business analytics in complex artificial intelligence environments. In *2019 IEEE International Conference on Fuzzy Systems* . IEEE.
5. Arora, M., Chopra, A. B., & Dixit, V. S. (2019). An approach to secure collaborative recommender system using artificial intelligence, deep learning, and blockchain. *Advances in Intelligent Systems and Computing*.
6. Ruiz-Real, J. L., Uribe-Toril, J., Torres, J. A., & De Pablo, J. (2020). Artificial intelligence in business and economics research: Trends and future. *Journal of Business Economics and Management*.
7. Caruso, S., Bruccoleri, M., Pietrosi, A., & Scaccianoce, A. (2023). Artificial intelligence to counteract "KPI overload" in business process monitoring: The case of anti-corruption in public organizations. *Business Process Management Journal*.
8. Jansen, J. A., Manukyan, A., Khoury, N. A., & Altuna Akalin. (2023). Leveraging large language models for data analysis automation. BioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2023.12.11.571140
9. Nasseri, M., Brandtner, P., Zimmermann, R., Taha Falatouri, Farzaneh Darbanian, & Tobechi Obinwanne. (2023). Applications of Large Language Models (LLMs) in Business Analytics – Exemplary Use Cases in Data Preparation Tasks. Lecture Notes in Computer Science, 182–198. https://doi.org/10.1007/978-3-031-48057-7_12
10. Chugh, T., Tyagi, K., Seth, R., & Srinivasan, P. (2023). Intelligent agents driven data analytics using Large Language Models. 30, 152–157. https://doi.org/10.1109/icoabcd59879.2023.10390973
11. Liu, S., Zhao, S., Jia, C., Zhuang, X., Long, Z., Zhou, J., Zhou, A., Lan, M., Wu, Q., & Yang, C. (2024). FinDABench: Benchmarking Financial Data Analysis Ability of Large Language Models.
12. Aggarwal, S., & Aggarwal, S. (2017). Deep investment in financial markets using deep learning models. *International Journal of Computer Applications*.
13. Aggarwal, S. (2017). Comparative analysis of hedge funds in financial markets using machine learning models. *International Journal of Computer Applications*.

14. Jaisinghani, B., & Aggarwal, S. (2024). adventureworks_2022_denormalized.csv (Version 1). *figshare*. https://doi.org/10.6084/m9.figshare.27899706.v1
15. bhave-sh. (2024). GitHub - bhave-sh/evaluating-genbi-paper: Source Files for Evaluation GenBI Article. GitHub. https://github.com/bhave-sh/evaluating-genbi-paper
16. Chainlit/chainlit. (2024, April 27). GitHub. https://github.com/Chainlit/chainlit
17. Introduction to PandasAI. (n.d.). PandasAI. https://docs.pandas-ai.com/intro