

Supplement to ‘Data Analytics with Large Language Models (LLM): A Novel Prompting Framework’

Shamma Mubarak Aylan Abdulla Almheiri, Mohammad AlAnsari,
Jaber AlHashmi, Noha Abdalmajeed,
Muhammed Jalil, and Gurdal Ertek

¹ College of Business and Economics, United Arab Emirates University, Al Ain, UAE
201003124@uaeu.ac.ae, 700040492@uaeu.ac.ae,
700040210@uaeu.ac.ae, 200935171@uaeu.ac.ae,
muhammad.jalil@uaeu.ac.ae, gurdal@uaeu.ac.ae

Abstract. This study presents a novel framework for conducting data analytics using Large Language Models (LLMs). The proposed framework suggests the construction of prompts and interaction patterns using four fundamental constructs: meta-specifications, specifications, instructions, and prompting patterns. The framework can guide and assist data engineers, analysts, and even non-technical domain experts by providing these four constructs as palettes of options. The LLM can then suggest analytics designs, conduct the analysis, provide posterior interpretations and insights, and produce other outputs, such as code or packaged software. The presented novel framework covers an immense space of possibilities through numerous combinations of selected meta-specifications, specifications, instructions, and prompting patterns. The primary theoretical contribution of this research is that it proposes a theoretical foundation and frame of reference for conducting data analytics using LLM. The primary practical contribution is that LLMs can now be employed much more systematically and extensively than before in designing and conducting data analytics. This opens a new world of applications powered by a countless combination of the four constructs across practically all fields of science, technology, and business, where LLMs can be used to guide, conduct, and interpret the results of data analytics.

Keywords: Data Analytics, Large Language Models, ChatGPT, Theoretical Framework, Bottom-Up Conceptual Analysis.

1 Introduction

Provided fully in the paper.

1.1 Preliminary Case Study

Provided fully in the paper.

1.2 Recent Developments

Since the preliminary case study was conducted in early 2023, until late 2023, when this research was completed, many developments have taken place in the LLM world. Some of these recent developments are as follows:

1. ChatGPT's "Advanced Data Analysis," formerly known as "Code Interpreter," has become an integral part of ChatGPT 4.0, directly accessible from within ChatGPT.
2. The code generation capability of ChatGPT, including the generation of Python, SQL, and R codes for data analytics, has improved considerably. Furthermore, many ChatGPT add-ins have been developed by third-party developers that are fine-tuned for code generation, informally referred to as "prompt-gramming," as opposed to the traditional "programming." One leading add-in is Grimoire [1], which was, as of January 2024, one of the top ChatGPT add-ins for code generation.
3. Several fine-tuned LLM models and platforms [2] that compete with ChatGPT with respect to code generation have been developed.
4. Popular spreadsheet desktops and cloud applications have evolved to either offer LLM capabilities or integrate with LLMs through add-ins/extensions. For example, MS Excel offers "Analyze Data" functionality directly from within MS Excel, which enables analyses such as Rank, Trend, Outlier, and Majority [3] [4]. With respect to integrations, the GPT for Work™ extensions [5] for Google Sheets [6] enables direct access to GPT and other LLM models from Google Sheets. For example, one can apply an analytics function or process multiple times through concatenated text prompts that read different source data/prompt cells each time.
5. Another important development is the emergence of a new generation of spreadsheets and analytics platforms built on generative AI. Two examples of such a platform are Numerous.ai [7] and Einblick [8].
6. Leading analytics platforms have readily integrated generative AI into various stages of analytics, including data cleaning, engineering, analysis, insight generation, and reporting. For example, the Tableau [9] visual analytics platform has integrated the AI capabilities of data storytelling, diagnostic insight derivation, and advanced predictive analytics democratized through no-code interfaces and wizards [10]. Another example of generative AI integration with analytics platforms is the online yEd Live graph analytics service [11], which can generate diagrams/graphs/networks as outputs for text prompts [12].
7. The leading mathematical and scientific computation platforms that have traditionally been used in research have introduced several novel solutions that incorporate LLMs. One of these solutions is LLMFunctions [13] by Wolfram [14], which extends the Wolfram language with LLM capabilities including symbolic chat, programmatic access to LLM functionality, content generation, and infrastructure functions. A more recent solution by Wolfram is Wolfram GPT [15], a ChatGPT plugin that enhances ChatGPT through advanced mathematical analysis, computation, curated knowledge/content creation, and analytics.

In light of recent developments and the explosion of innovations, as a novel stage and unique contribution, a framework was developed in this research study for conducting data analytics with LLM, through Bottom-Up Conceptual Analysis [16].

1.3 Framework Proposition

Provided fully in the paper.

1.4 Unique Contributions

Provided fully in the paper.

2 Literature

Provided fully in the paper.

2.1 LLM

Provided fully in the paper.

2.2 LLM for Data Analytics

The research literature has a growing number of studies have reported the application of LLMs in data analytics. Most of the studies were published as preprints under Arxiv.org [17]. The applications cover diverse areas of science, technology, business, and humanities, including conducting data analytics in precision agriculture [18], medical data augmentation through medication identification and medical event classification [19], detecting software vulnerabilities [20], supporting data analytics processes for telematics data analysis [21], and exploring shifts in cultural values [22].

[23] assessed the data analysis and visualization capabilities of various LLMs with respect to three criteria, namely correctness, efficiency, and comprehensibility, suggesting that GPT LLM models perform better than others.

Generative AI can be used not only to conduct specific tasks and steps of data analytics but also to generate complete data analytics/data science workflows [24].

2.3 Constructs of LLM

The effective use of LLMs in data analytics requires careful consideration of several aspects. Below, we summarize these aspects in terms of the techniques and LLM parameters that should be carefully configured for optimal LLM performance. The structures of various LLMs are visually documented in [25].

Techniques

Transfer Learning: LLMs employ transfer learning, where models pretrained on massive datasets can be fine-tuned for specific tasks or domains. This approach allows efficient knowledge transfer and adaptation to various data analytics applications [26].

Attention Mechanism: The attention mechanism in LLMs facilitates the model's ability to focus on specific parts of the input sequence, thereby enhancing its contextual understanding. This mechanism is pivotal for capturing relationships and dependencies within data [27].

Parameters

Model Size: The size of an LLM, often measured by the number of parameters, significantly influences its performance. Larger models tend to capture more nuanced patterns but may require substantial computational resources [26].

Learning Rate and Batch Size: Configuring parameters such as learning rate and batch size during fine-tuning is crucial for achieving optimal performance in specific data analytics tasks. Careful tuning helps balance model convergence and training efficiency [28].

2.4 Prompt Engineering

Provided fully in the paper.

2.5 Summary

Provided fully in the paper.

3 Preliminary Case Study

Provided fully in the paper.

3.1 Interaction with the LLM

The steps of interaction with the LLM are provided in the paper. The user prompts during the interaction with the LLM are provided in Table 1 in this document.

3.2 LLM Outputs

Provided fully in the paper.

3.3 Observations from the Case Study

Provided fully in the paper.

Table 1. The prompts for the first seven steps of the interaction with LLM in the case study.

Name	Description
Step 1	Hi ChatGPT, I have a sales transactions dataset consisting of the following attributes, with description each attribute explained next to it, basically in the following format: Variable: Description What I want from you is to list for me a list of top 10 pivot tables that I might construct with this data, that can give me and a sales manager the most insights. Can you do that?
Step 2	Here are my attributes and their explanations: <ul style="list-style-type: none"> - Inv_Number: Inventory number of the item sold. - Store_Num: Store number where this item line was sold. - Description: Description of the item sold. - Price: Price of the item sold. - Sold: Quantity sold for this observation; negatives indicate sell backs. - Del: Deliveries entered (denominated in std count units) plus Sum of Transfers In (+) and Transfers Out (-). - Sales: Sales in dollars for this observation. - Tot_Sls: The percentage of total sales for this week at this store that the item makes up. - Unit_Cost: Cost per unit for the specified inventory item at the specified store. - Cost: Total cost for the observation. - Cost_Percent: Total sales divided by cost. - Margin: Profit margin for this observation. - Profit: Total gross profit for this observation. - Year: Year the item was sold. - Month: Month the item was sold. - Day: Day of the month for the week's observation.
Step 3	Thank you so much. This is brilliant. How did you decide that these are the top 10 analyses that should be done, before any other pivot table? If I used the pivot table designs that you suggested, how can I convince my client that these are indeed the best analyses to be prioritized?
Step 4	Can you provide for me any statistics, such as number of search results in google, or supporting evidence such as an academic paper or a business whitepaper, that can support the choice of each pivot table design?
Step 5	For each of these pivot table designs, I also want to visualize the results of that pivot table using the most appropriate data visualization chart/plot. So, could you please recommend for me the most suitable type of chart/plot for each design, and also tell me how I should construct that design that you are suggesting. For example, if you suggest me a line chart, please also tell me what should go on the X axis, Y axis, color of the lines, size of the lines, and other attributes of the visual elements. Can you suggest now charts/plots for the top 10 pivot tables you suggested, together with their design, as well?
Step 6	Thank you so much. This is brilliant. How did you decide that these are best charts/plots that should be constructed, before any other types of chart/plot? If I used the chart/plot designs that you suggested, how can I convince my client that these are indeed the best chart/plots to be prioritized?
Step 7	Can you provide for me any statistics, such as number of search results in google, or supporting evidence such as an academic paper or a business whitepaper, that can support the choice of each chart/plot design?

4 Proposed Framework

Provided fully in the paper.

4.1 Meta-Specifications

Provided fully in the paper.

4.2 Specifications

Provided fully in the paper.

4.3 Instructions

Provided fully in the paper.

4.4 Prompt Engineering Patterns

Provided fully in the paper.

5 Conclusion

Provided fully in the paper.

Acknowledgement

This research was funded by the “CBE Annual Research Program (CARP)” for the academic year 2023-2024. The funding is provided by the College of Business and Economics (CBE) of the United Arab Emirates University (UAEU).

References

1. Mind Goblin Studios. <https://gptavern.mindgoblinstudios.com/>, last accessed 2024/01/12.
2. Zheng, Z., Ning, K., Wang, Y., Zhang, J., Zheng, D., Ye, M., & Chen, J. A Survey of Large Language Models for Code: Evolution, Benchmarking, and Future Trends (2024). arXiv preprint arXiv:2311.10372.
3. Microsoft. <http://tinyurl.com/36km2fjp>, last accessed 2024/01/07.
4. YouTube, Leila Garani. How to use Analyze Data in Excel (AI Creates Pivot Tables and Charts). <https://www.youtube.com/watch?v=AXNPR5q1y08>, last accessed 2024/01/07.
5. GPT for Work. <https://gptforwork.com/>, last accessed 2024/01/07.
6. Google Sheets. <https://docs.google.com/spreadsheets/>, last accessed 2024/01/07.
7. Numerous. <https://numerous.ai/>, last accessed 2024/01/13.
8. Einblick. <https://www.einblick.ai/>, last accessed 2024/01/14.
9. Tableau. <https://www.tableau.com/>, last accessed 2024/01/07.
10. Tableau AI. <https://www.tableau.com/products/tableau-ai>, last accessed 2024/01/07.

11. yEd Live. <http://www.yworks.com/products/yed-live>, last accessed 2024/01/07.
12. yWorks. ChatGPT and yEd Live (2023). <http://www.yworks.com/blog/chatgpt-and-yed-live>, last accessed 2024/01/07.
13. Wolfram Cloud. Wolfram/LLMFunctions. Language model and other API based machine learning functions for the WL. <https://resources.wolframcloud.com/PacletRepository/resources/Wolfram/LLMFunctions/> last accessed 2024/01/12.
14. Wolfram. <https://www.wolfram.com/>, last accessed 2024/01/12.
15. Wolfram GPT. <https://gpt.wolfram.com/>, last accessed 2024/01/12.
16. Salmon, W. *Scientific Explanation and the Causal Structure of the World*, Princeton, N.J.: Princeton University Press (1984).
17. Arxiv.org. <https://arxiv.org/>, last accessed 2024/01/14.
18. Potamitis, I. ChatGPT in the context of precision agriculture data analytics. arXiv preprint arXiv:2311.06390 (2023). <https://arxiv.org/ftp/arxiv/papers/2311/2311.06390.pdf>
19. Sarker, S., Qian, L., Dong, X. Medical Data Augmentation via ChatGPT: A Case Study on Medication Identification and Medication Event Classification. arXiv preprint arXiv:2306.07297 (2023). <https://arxiv.org/pdf/2306.07297.pdf>
20. Fu, M., Tantithamthavorn, C., Nguyen, V., Le, T. Chatgpt for vulnerability detection, classification, and repair: How far are we?. arXiv preprint arXiv:2310.09810 (2023). <https://arxiv.org/pdf/2310.09810.pdf>
21. Lingo, R. The Role of ChatGPT in Democratizing Data Science: An Exploration of AI-facilitated Data Analysis in Telematics. arXiv preprint arXiv:2308.02045 (2023). <https://arxiv.org/pdf/2308.02045.pdf>
22. Vargas-Solar, G., Cerquitelli, T., Espinosa-Oviedo, J.A., Cheval, F., Buchaille, A., Polgar, L. Conversational Data Exploration: A Game-Changer for Designing Data Science Pipelines. arXiv e-prints, arXiv-2311 (2023). <https://arxiv.org/pdf/2311.06695.pdf>
23. Nejjar, M., Zacharias, L., Stiehle, F., & Weber, I. LLMs for Science: Usage for Code Generation and Data Analysis. arXiv preprint arXiv:2311.16733 (2023). <https://arxiv.org/pdf/2311.16733.pdf>
24. Duque, A., Syed, A., Day, K.V., Berry, M.J., Katz, D.S., Kindratenko, V.V. Leveraging Large Language Models to Build and Execute Computational Workflows. arXiv preprint arXiv:2312.07711 (2023). <https://arxiv.org/pdf/2312.07711.pdf>
25. LLM Visualization. <https://bbycroft.net/llm>, last accessed 2024/01/14.
26. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ..., Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, MF & Lin, H.) 33 (Curran Associates, Inc., 2020), 1877–1901 (2020). arXiv preprint arXiv:2005.14165.
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ..., Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, 30 (2017). <http://tinyurl.com/mntyxtc9>, last accessed 2024/01/14.
28. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.