Large Language Models for Intelligent Coaching in Data Science Problem Solving: A Preliminary Investigation

Maryam Alomair[1,2,†], Shimei Pan[1] and Lujie Karen Chen[1]

[1]University of Maryland Baltimore County, Baltimore, MD, USA 21250
[2]King Faisal University, Ahsaa, KSA 31982

Abstract

Data Science Problem Solving (DSPS) is a complex skill set involving domain knowledge, analytical methods, and critical thinking to formulate and refine problem statements, analyze data, apply appropriate methods, diagnose model performance, interpret results, and communicate findings to stakeholders. DSPS aligns with a high level of Bloom's Taxonomy of applying, analyzing, and evaluating. With the increasing power of artificial intelligence (AI) tools like ChatGPT to support lower-level cognitive tasks, such as writing codes, proficiency in advanced skills like DSPS is rising. Due to the complexity of DSPS, efficient and scalable approaches to support learning are underdeveloped. To explore the potential of using Large Language Models (LLMs) as coaching assistants for DSPS, we evaluated three popular LLMs: GPT-3.5 (a.k.a. ChatGPT), GPT-4, and Google Bard, using a set of case-based DSPS problem sets. This evaluation focuses on two critical dimensions: (I) the accuracy of LLMs' responses to DSPS questions and (II) their capacity to provide high-quality explanations. The human expert evaluation demonstrates the promise of LLMs, especially GPT-4, in generating accurate answers and high-quality explanations. We discuss the implication of the results in developing novel LLM-based tools to facilitate large-scale individualized DSPS learning.

Keywords

Data Science Problem Solving, Large Language Models, Explanation

1. Introduction

With the rise of Large Language Models (LLMs), more data science tasks are automatable, shifting the role of data scientists towards higher-order thinking [1]. LLMs also enable personalized student support, revolutionizing data science education in content and pedagogy [1]. Our paper addresses LLMs' impact on education, focusing on coaching students in data science problem solving (DSPS), which is a crucial skill that is less prone to automation.

DSPS encompasses complex skills required for applying appropriate data science techniques to solve real-world problems. It builds on the foundational conceptual knowledge of data science methods and the procedural knowledge of coding to implement these methods. DSPS represents a form of "conditional knowledge" [2], which involves understanding when and why to use specific methods, aligning with the higher-order cognitive processes described in Bloom's Taxonomy, specifically the creation level. Developing DSPS requires a cognitive approach distinct from that for lower-level knowledge and skills, often honed through practical experiences, such as internships or mentored projects. However, the data science curriculum offers limited opportunities for large-scale practice of these skills due to its labor-intensive nature. The Intelligent Tutoring System (ITS) provides a framework to address this challenge. Our vision entails an AI-powered coaching assistant acting as a conversational virtual mentor in students' learning environments, aiding them in solving real-world problems. This system will incorporate a **domain model** of DSPS, maintain a detailed **student model** tracking mastery of DSPS skills, and utilize a **pedagogical model** offering timely coaching and resource recommendations. Generative AI is poised to play a crucial role in these models.

In this paper, we evaluate LLM's implicit domain model of DSPS, a foundational component of ITS, by analyzing its ability to provide accurate responses and reasonable explanations to a case-based DSPS problem set called Caselet. Developed over several years by data science experts, Caselet is designed to scaffold students' DSPS skills through bite-sized case studies. To our knowledge, this is the first systematic, targeted evaluation of LLMs' problem-solving capacity, compared with most other works that are either based on case studies [1] or focus on well-specified analyst tasks [3]. Overall, our research addresses the gap in data science education research concerning higher-order competencies and contributes to the emerging field of leveraging AI to enhance access to high-quality data science training opportunities at a large scale.

2. Related Work

Data science combines computing with mathematical or statistical reasoning. Although a precise definition of data science competence is still developing, it is widely accepted that data scientists must write code for various analytical or modeling procedures, a skill closely related to programming. The assessment of LLMs' programming abilities has generated considerable interest. One study [4] explores Codex, a deep learning model trained on Python code, and its proficiency in handling introductory programming tasks, where it outperforms most students in real exam scenarios. Another study [5] shows that GitHub Copilot can solve about half of the problems on the first try and improves to solve 60% of the rest with iterative adjustments. ChatGPT has also been used to effectively support undergraduate Computer Science students with programming challenges [6], despite some inaccuracies in the code. Furthermore, the role of LLMs in providing code explanations has been examined [7], with findings indicating that students generally find these explanations beneficial, although their experience varies with the complexity of the code, the type of explanation, and the length of the code snippet.

Educational Datamining '24 Human-Centric eXplainable AI in Education and Leveraging Large Language Models for Next-Generation Educational Technologies Workshop Joint Proceedings, July 13, 2024, Atlanta, GA
*Corresponding author.
maryama4@umbc.edu (M. Alomair); shimei@umbc.edu (S. Pan); lujiec@umbc.edu (L. K. Chen)
0009-0008-8343-5814 (M. Alomair); 0000-0002-5989-8543 (S. Pan); 0000-0002-7185-8405 (L. K. Chen)

The evaluation of LLMs' data science competency beyond coding, however, is quite limited so far. Tu et al. [1] presented a case study using ChatGPT and a code interpreter to execute a data science pipeline for a Kaggle dataset, covering data cleaning, exploration, model building, interpretation, and result presentation. This study also assessed ChatGPT's ability to answer statistical exam questions, achieving an impressive score of 104 out of 116. Cheng et al. [3] evaluated GPT-4's performance on well-specified data analyst tasks, including composing database queries, generating charts, and deriving insights. GPT-4's performance was comparable to that of senior data analysts. Additionally, Chen et al. [8] evaluated GPT-3.5 and GPT-4 on a diverse range of data visualization tasks, including data cleaning, data exploration, interactive visualization, and insights communication. This evaluation was based on assignments from a data visualization course at Harvard University and achieved an accuracy of 80%. Overall, systematic evaluation of LLMs in data science education, particularly in providing support for higher-order reasoning tasks like DSPS, remains relatively sparse, which is the focus of this paper.

3. Method

3.1. Caselet Practices for DSPS

Caselets [9] are a case-based practice tool we developed to support the development of DSPS competency for data science learners on a large scale. Each Caselet is a bite-sized case study that starts with a problem context connecting real-world application scenarios, followed by a data summary describing the property of the dataset(s). It then presents 5-7 multi-select multiple-choice questions (i.e., multiple correct answers are possible). Upon completion, learners will receive question-level feedback (correct or incorrect) and explanations. Caselets was written by a team of experienced data scientists who drew from their real-world data science problem-solving experience. Caselets' questions cover the main knowledge components (KCs) of DSPS. KCs are utilized to describe mental processes at a granularity level approximately corresponding to individual task units [10]. KCs for DSPS include Problem Formulation, Data Understanding, Data Preprocessing, Model Selection, Model Configuration, Experiment Design, and Model Diagnosis. Those KCs practiced through Caselets represent critical decision-making points in the DSPS process. We have compiled 10 Caselets with 70 questions, which have been piloted at two institutes and used by 183 graduate students as a part of their course assignments.

3.2. Prompt LLMs Using Caselet Questions

We used three LLMs in our experiments: GPT-3.5 (a.k.a. ChatGPT, a free version from Open AI), GPT-4 (a paid version from Open AI), and Google Bard (free from Google). They are versatile models, excelling in diverse natural language processing tasks with no or very limited explicit training [11] (This is commonly known as zero-shot or few-shot learning). Between GPT-3.5 and GPT-4, GPT-4 is a more advanced model with an order of magnitude larger than GPT-3.5. This enables the GPT-4 model to understand context and distinguish nuances better, resulting in more accurate and coherent responses [12]. Google Bard uses the Language Model for Dialogue Applications (LaMDA) architecture. Bard is trained on an extensive dataset encompassing both text and code, giving it a more extensive world understanding and the ability to generate comprehensive, informative responses. In contrast, GPTs trained solely on text often excel in producing creative and engaging responses [13]. To assess LLMs' capacity to provide reliable answers, we prompt LLMs in a zero-shot setting to answer and explain Caselet questions. We chose a zero-shot prompt to highlight the model's ability to generalize and provide coherent outputs in situations where it hasn't been specifically trained. Please refer to http://tinyurl.com /2jnjd75t for a prompt example. We use the same prompt structure for all Caselets and all LLMs. Given the current limitations of the LLMs in handling non-textual contents, we made the following adaptation in presenting Caselet questions. (1) Data summary presented in the format of PDF, HTML, or tables are omitted; (2) Questions that rely on the interpretation of plots are omitted. As a result, we used 9 Caselets in our study, among which two were presented without data summaries. Among the 52 questions, 39 were selected as they do not rely on plot or table comprehension. Among these questions, GPT-3.5 and GPT-4 generated responses to 38 questions, while Google Bard answered 36 questions.

3.3. Evaluate Accuracy of Responses and Quality of Explanations

Accuracy of Responses We assess the responses generated by LLMs based on their alignment with the correct answers. Responses that perfectly match the correct choices are categorized as "correct." Conversely, if there is no correspondence between the response and the correct options, it is categorized as "wrong." Given that questions can have multiple correct choices, LLMs sometimes select some correct choices while omitting others or blending correct and incorrect choices. In such cases, we designate the responses as "partially correct." In addition to evaluating overall accuracy, we also conducted a detailed analysis of response accuracy in relation to KCs. **Quality of Explanation** We evaluated the capability of LLMs in generating high-quality explanations using a subset of 13 questions that were correctly answered by all three models, compared with explanations authored and validated by human experts. To ensure the rigor of our evaluation, we asked three experienced data scientists as evaluators, with a minimum of three years of experience in data science. They were tasked with rating the quality of the explanations generated by the human experts and the LLMs for each of the 13 questions, with 52 explanations in total. To maintain impartiality, our evaluators were blinded to the author of each explanation. The evaluators assessed each explanation using a rating scale ranging from 1 (low quality) to 4 (high quality), using a rubric of nine criteria established in existing explanation literature [14, 15], including accuracy, completeness, readability, relevance, use of analogy, definition of concepts, use of everyday language, inclusion of examples, and storytelling. Detailed information on this rubric can be found at http://tinyurl.com/2jnjd75t. The inter-rater agreement among the evaluators is between 0.51 and 0.67 (Fleiss' kappa score), indicating a moderate to good level of agreement among the evaluators.

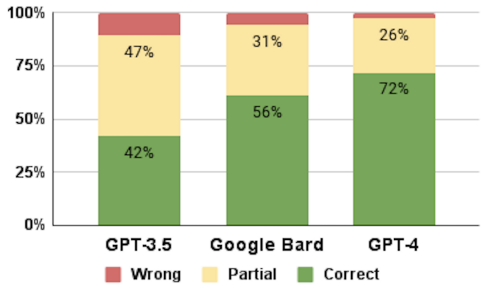

Figure 1: The overall accuracy rate, evaluated on three LLMs.

		Accuracy (%)		
Knowledge Component	**N**	**GPT-3.5**	**Google Bard**	**GPT-4**
Problem Formulation	10	44	67	90
Data Processing	4	33	67	50
Data Understanding	3	50	100	100
Model Selection	6	40	83	83
Model configuration	3	50	33	100
Experiment Design	3	0	33	33
Model Diagnosis	10	33	44	50

Figure 2: The complete accuracy rate by data science knowledge components, evaluated on three LLMs.

4. Results

4.1. Accuracy of Responses

Figure 1 summarizes the **overall accuracy** of the three LLMs. The results show that GPT-4 outperformed the other two models in providing completely correct answers at a frequency of 72%, compared to 56% for Google Bard and 42% for GPT-3.5. We also noted that GPT-4 and Bard rarely generated completely wrong answers (3% for GPT-4 and 5% for Bard), compared to GPT-3.5's 11%. Out of the 39 questions, only a single question remained unanswered by all the three LLMs. Furthermore, Bard could not answer three questions, attributing this to its status as a language model. However, GPT-4 generated responses for these questions, with one being completely accurate and the other being partially correct. GPT-3.5 couldn't answer one of the three questions, attributing this to limitations in its AI capabilities. To further understand LLMs' capacity to provide reliable answers to questions on different KCs, we analyzed the **rate of complete correctness by KCs**. As summarized in Figure 2, we note consistent patterns among the three models across all KCs, with GPT-4 outperforming (at least no worse than) Google Bard, which in turn has an edge over GPT-3.5. Focusing on the best performing GPT-4 on the two most frequent KCs, it is interesting to note that GPT-4 can achieve an impressive 90% accuracy for the problem formulation KC while less reliable for the model diagnosis KC (50%). Bard was less reliable for the problem formulation KC (67%). GPT-3.5 performed the worst on the problem formulation KC (44%). All three models seem to struggle with the model diagnosis KC, with their accuracy no better than 50%.

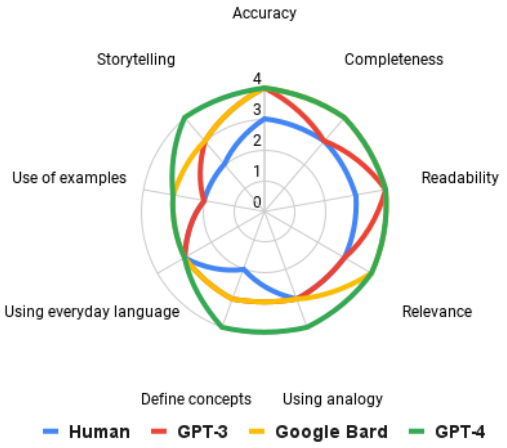

Figure 3: Human evaluations of multi-dimensional quality of explanations generated by human experts and LLMs.

4.2. Quality of Explanation

As illustrated in Figure 3, GPT-4 emerges as the best performer, consistently receiving the highest scores in 7 of the criteria. This highlights its potential to produce comprehensive, engaging, and informative explanations. Across the board, all generators scored 3 out of 4 in the "Using everyday language" criterion. GPT-3.5 and Bard demonstrated reasonable competitiveness in several criteria. Interestingly, the rubric criteria led evaluators to categorize human explanations as less compelling with respect to "Defining Concepts", "Using Examples" and "Storytelling".

We also conducted a qualitative analysis to identify patterns in how explanations are given from a pedagogical point of view. We note that the explanations seem to follow a predictable structure, including a restatement of the question, the correct answer, and the reason for choosing the correct answer and not choosing the incorrect answer. It's interesting to note that when they introduce a new term, like "overfitting" or "regularization," that was not mentioned in the question or answer options, they usually give a short definition of the term. In some instances, we found that the language model can provide additional insights based on common knowledge. For example, when discussing the factors to be considered for a house price model, GPT-4 explained: "Changes in the overall economy can have a major impact on home prices. For example, during a recession, home prices might fall due to decreased demand." This extra information could make the explanation more accessible to learners.

5. Conclusion and Future Work

In this study, we explored the potential of using LLMs as intelligent coaching assistants for DSPS, focusing on their ability to provide accurate responses and high-quality explanations. We evaluated three popular LLMs: GPT-3.5 (a.k.a. ChatGPT), GPT-4, and Google Bard, using a set of case-based DSPS problem sets called Caselets. Human evaluation and quantitative measures have demonstrated the promise of LLMs, especially GPT-4, in generating accurate answers and high-quality explanations. We also note the varied capacities of LLMs in providing correct answers to questions focusing on different types of KCs. Specifically, we note

that LLMs perform better on the problem formulation KC and struggle with the model diagnosis KC. One possible explanation for the heterogeneity in LLMs' performance by KC is that problem formulation KC relies on reasoning based on textual information, while model diagnosis KC requires reasoning over quantitative information (e.g., model performance), which seems to be a weakness for LLMs in general.

It is worth noting that though there were cases where the LLMs excelled, there were also situations where they answered questions partially or entirely wrong. These incorrect responses, however, were delivered with apparent confidence. This poses a risk for beginners who need help distinguishing between correct and misleading explanations. It emphasizes the need for transparency: learners should be informed that specific explanations come from LLMs, not humans, and be alerted to potential errors despite their authoritative tones [16]. Engaging critically with LLM-generated instruction content should be an integral part of data science education in the era of generative AI. As LLMs continue to evolve and improve, repeating these experiments to guide their future development is essential. To improve upon the baseline that relies on zero-shot learning, we plan to investigate the usefulness of LLMs in assisting DSPS in few-shot learning settings. Moreover, we will also explore the impact of model temperature on LLMs' ability to generate accurate answers and high-quality explanations.

Our study offers insights into using AI for data science education, focusing on LLMs within the Intelligent Tutoring System framework. We demonstrate LLMs' understanding of DSPS and their potential to generate effective explanations. In the future, we will integrate student models and explore human coaches' roles in AI-assisted learning to maximize an intelligent coaching system's potential.

References

[1] X. Tu, J. Zou, W. Su, L. Zhang, What should data science education do with large language models?, Harvard Data Science Review (2024).

[2] P. H. Winne, R. Azevedo, Metacognition, Cambridge Handbooks in Psychology, Cambridge University Press, 2014, p. 63–87.

[3] L. Cheng, X. Li, L. Bing, Is gpt-4 a good data analyst?, arXiv preprint arXiv:2305.15038 (2023).

[4] J. Finnie-Ansley, P. Denny, B. A. Becker, A. Luxton-Reilly, J. Prather, The robots are coming: Exploring the implications of openai codex on introductory programming, in: Proceedings of the 24th Australasian Computing Education Conference, ACE '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 10–19. URL: https://doi.org/10.1145/3511861.3511863. doi:10.1145/3511861.3511863.

[5] P. Denny, V. Kumar, N. Giacaman, Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language, in: Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2023, Association for Computing Machinery, New York, NY, USA, 2023, p. 1136–1142. URL: https://doi.org/10.1145/3545945.3569823. doi:10.1145/3545945.3569823.

[6] B. Qureshi, Exploring the use of chatgpt as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges, 2023. arXiv:2304.11214.

[7] S. MacNeil, A. Tran, A. Hellas, J. Kim, S. Sarsa, P. Denny, S. Bernstein, J. Leinonen, Experiences from using code explanations generated by large language models in a web software development e-book, 2022. arXiv:2211.02265.

[8] Z. Chen, C. Zhang, Q. Wang, J. Troidl, S. Warchol, J. Beyer, N. Gehlenborg, H. Pfister, Beyond generating code: Evaluating gpt on a data visualization course, arXiv preprint arXiv:2306.02914 (2023).

[9] L. Chen, A. Dubrawski, Accelerated apprenticeship: teaching data science problem solving skills at scale, in: Proceedings of the Fifth Annual ACM Conference on Learning at Scale, 2018, pp. 1–4.

[10] K. R. Koedinger, A. T. Corbett, C. Perfetti, The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning, Cognitive Science 36 (2012) 757–798.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[12] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. doi:10.1016/j.iotcps.2023.04.003.

[13] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects, 2023. URL: https://doi.org/10.36227/techrxiv.23589741.v3. doi:10.36227/techrxiv.23589741.v3, preprint.

[14] C. Kulgemeyer, Towards a framework for effective instructional explanations in science teaching, Studies in Science Education 54 (2018) 109–139.

[15] J. Wittwer, A. Renkl, Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations, Educational Psychologist 43 (2008) 49–64.

[16] NIST, AI Risk Management Framework, 2021. URL: https://www.nist.gov/itl/ai-risk-management-framework.