JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 12(1), 65–90. https://doi.org/10.18608/jla.2025.8615

Exploring the Potential of Generative AI to Support Non-experts in Learning Analytics Practice

Xavier Ochoa[1]*, Xiaomeng Huang[2], Yuli Shao[3]

Abstract

Generative AI (GenAI) has the potential to revolutionize the analysis of educational data, significantly impacting learning analytics (LA). This study explores the capability of non-experts, including administrators, instructors, and students, to effectively use GenAI for descriptive LA tasks without requiring specialized knowledge in data processing, visualization, or programming. Through a laboratory experiment, participants with varying levels of expertise in data analysis engaged in three tasks with different levels of difficulty using ChatGPT. The findings reveal that while there is a small effect of previous expertise on performance, novices and experts achieved remarkably similar scores. Additionally, the study identifies that action sequence variables, such as the sequence's complexity and the presence of specific actions such as evaluating and checking results, significantly predict performance. These results suggest that while current GenAI technologies are not yet ready to fully support non-experts, they hold the promise of supporting stakeholders, regardless of their technical background, to perform descriptive data analysis in the context of LA practice. This research seeks to start a discussion within the LA community about leveraging AI to scale and expand LA practices, potentially transforming how educational stakeholders engage with and benefit from LA.

Notes for Practice

- Generative AI (GenAI) technologies, like ChatGPT, have the potential to significantly lower the barriers for non-experts to engage in learning analytics tasks. This enables a broader range of stakeholders, including educators without technical expertise, to conduct sophisticated data analyses.
- While GenAI can assist in the technical aspects of data analysis, it does not replace the need for domain knowledge. Practitioners should have a solid understanding of the educational context to effectively interpret, check, and evaluate the AI-generated results.
- Novices achieved good scores using current GenAI tools, but the results are not perfect. These tools, in their current state, do not yet support completely independent use for solving real-life problems.

Keywords

GenAI, scaling learning analytics, data analysis expertise.

Submitted: 08/08/2024 — **Accepted:** 27/01/2025 — **Published:** 15/03/2025

[1] *Corresponding author Email: xavier.ochoa@nyu.edu Address: Steinhardt School of Culture, Education and Human Development, New York University, 370 Jay St. Brooklyn, New York, USA. ORCID iD: https://orcid.org/0000-0002-4371-7701

[2] Email: xiaomeng.huang@nyu.edu Address: Steinhardt School of Culture, Education and Human Development, New York University, 370 Jay St. Brooklyn, New York, USA. ORCID iD: https://orcid.org/0000-0002-6992-061X

[3] Email: ys3203@nyu.edu Address: Steinhardt School of Culture, Education and Human Development, New York University, 370 Jay St. Brooklyn, New York, USA.ORCID iD: https://orcid.org/0009-0002-3726-0677

1. Introduction

AI is not a new field (Toosi et al., 2021). It underpins many technologies, from sophisticated applications like fraud detection in credit card transactions (Plakandaras et al., 2022) and medical image diagnostics (Castiglioni et al., 2021) to simpler ones like detecting smiles in digital cameras (Goswami et al., 2021) and managing washing machine cycles (Raja & Ramathilagam, 2021). Historically, using AI algorithms (e.g., support-vector machines, fuzzy logic) required specialized expertise, limiting their application to scientists and engineers. This paradigm changed in 2022 when OpenAI released ChatGPT, a user-friendly chat application that lets users interact with GPTv3, a transformer-based large language model (LLM) trained on vast textual data. While building AI tools remains the realm of experts, using cutting-edge AI tools to solve problems—from sophisticated to mundane—became as simple as clearly communicating the problem and expected solution in plain language.

This democratization of AI parallels the advent of personal computers in the early 1980s (Ceruzzi, 1996) and the introduction of home Internet in the late 1990s (Leiner et al., 2009). Both technologies existed before becoming widely accessible, but their use as problem-solving tools was limited to a select few intermediaries. These past revolutions gave most people access to computing power and communication tools. Now, the AI revolution, driven by generative AI (GenAI) models and services (Stokel-Walker & Van Noorden, 2023), is providing human-level problem-solving and creative skills to the masses.

Learning analytics (LA), with its tradition of using AI to analyze and model learners' actions and mental states, has rapidly integrated these new tools. This is evident from numerous works presented at the Learning Analytics and Knowledge 2024 conference (Flanagan et al., 2024) and the 2023 editorial in the *Journal of Learning Analytics* (Khosravi et al., 2023). The most significant publication to date is a position paper by Yan and colleagues (2024), which explores how GenAI can transform LA research and practice. They propose six research directions: (1) understanding hybrid human-AI learners, (2) using AI to understand unstructured data, (3) enhancing visual and explanatory analytics, (4) personalizing analytic information, (5) enhancing accessibility with AI, and (6) ethical AI integration. While all are important, we focus on directions 3, 4, and 5, which relate to facilitating access to LA practice.

Direction 3 invites researchers to move beyond traditional expert-designed dashboards (Verbert et al., 2020) and leverage the conversational capabilities of LLMs to create more user-friendly methods for interacting with LA information. This could include contextual explanations and data storytelling (Yan et al., 2024). The expectation is that these new interfaces will reduce the need for students and teachers to interpret complex visualizations (Martinez-Maldonado et al., 2020), thereby lowering barriers to using and reflecting on LA insights. Direction 4 suggests using LLMs to tailor LA information or feedback to the specific needs of different users at various stages of their learning journey (Yan et al., 2024). One way to achieve this is by allowing LLMs to dynamically respond to ad hoc queries from individual users without relying on predefined dashboards or use cases (Zhu et al., 2024). This approach would eliminate the need for purpose-built interfaces, replacing them with a flexible, adaptable, and easily expandable conversational interface. Finally, direction 5 makes a call to ensure that advancements in LA using LLMs are accessible to all learners (Yan et al., 2024). While Yan and colleagues (2024) focus on socio-economic aspects, we interpret this call as extending the benefits of LA to everyone, regardless of their data literacy level, access to experts, or technical skills.

Specifically, we are interested in exploring how GenAI could facilitate LA practices by moving beyond predefined, expert-built dashboard interfaces. This would involve introducing LLM-powered conversational interfaces capable of providing personalized and contextualized data visuals, along with verbal explanations, enabling stakeholders to address their individual questions without requiring technical expertise in data analysis.

To understand the potential impact of these tools on LA practices, we should revisit the LA cycle (Clow, 2012). According to this cycle, the analytic process begins with learners engaging in actions during their learning journey. Data about these actions—referred to as learning traces—is collected and stored. This data then needs to be selected, cleaned, wrangled, and analyzed to produce metrics, indicators, or feedback. Finally, this information is presented back to stakeholders in the learning process, such as teachers, advisors, or the learners themselves, through an LA intervention (Wise, 2014). Ideally, this intervention triggers positive changes in the learning process, which are reflected in new or altered actions by the learner, thus continuing the cycle.

From this high-level overview of the LA process, it becomes clear that two of these steps, due to their technical complexity, are beyond the reach of most LA stakeholders, who do not possess a background in data analysis (they will be called non-experts in the remainder of this work). The first one of these steps is data collection, which often involves the construction and deployment of physical or virtual sensors to capture learners' actions (Ochoa, 2022). The second is data processing, analysis, and visualization. This step requires expertise in data science and programming to design and develop LA tools, typically in the form of dashboards, that present the data to stakeholders (Klašnja-Milićević et al., 2017). While reliably assisting stakeholders in building software or hardware sensors to capture and store learning traces is still beyond the current capabilities of GenAI, we propose that leveraging LLMs to support stakeholders in conducting their own data analysis and visualization not only is feasible but could have the most significant immediate impact on transforming LA practices.

As will be discussed in Section 2.3, initial experimentation in the field of data science (Zheng, 2023; Shen et al., 2024) provides credible evidence that conversational interfaces powered by LLMs can automate data processing, analysis, and visualization. In these interfaces, users can provide data in various formats and pose ad hoc questions. The LLM generates code that is executed automatically by a code interpreter, and the results are presented to the user in the form of tables or visualizations or are used by the LLM to prepare a verbal interpretation. However, these studies primarily focus on AI use by experts or data science students and do not provide evidence on how effective current GenAI technologies are in supporting individuals with varying levels of data analysis expertise to obtain accurate answers independently. Furthermore, these studies do not address the most common type of data analysis used in LA dashboards (LADs), namely, descriptive data analysis (Paulsen & Lindsay, 2024). Additionally, the use of GenAI to assist non-experts in conducting descriptive LA raises several more nuanced questions beyond simple performance: What level of data analysis expertise is still required from stakeholders?

How complex can their questions be? Can stakeholders detect and correct errors in the analysis? What processes do they follow?

The purpose of the present work is to obtain initial answers to these questions. By gaining insights into how current stakeholders, especially teachers and instructors, could use current GenAI technology, the field could start designing, implementing, and evaluating new ways to conduct LA practice, in which the analytic process is guided and executed by those stakeholders without the need to depend on intermediary experts to help them answer their questions. Also, by understanding the current limitations and points of improvement, the field could play a role in shaping the evolution of GenAI models more adapted for LA and better prepare educational stakeholders to benefit from these models.

To obtain answers to these questions, this work, in Section 2, first examines what prevents non-experts from conducting descriptive analysis, the current ways in which descriptive analytics are conducted in LA, and what have been the experiences of using GenAI for general data analysis in adjacent fields. Then, in Section 3, this paper describes the experimental design used to evaluate how well current state-of-the-art GenAI systems can support individuals with varying levels of expertise in data analysis to construct descriptive LA questions of diverse complexity. Section 4 analyzes the experimental data and presents the results of those analyses. Section 5 discusses the implications of these results for the LA research and practice communities. Finally, Section 6 outlines the limitations of the study and suggests directions for future work.

2. Literature Review

Descriptive analytics involves analyzing historical data to summarize, interpret, and understand past events. It answers the question "What happened?" by providing insights into past performance through data filtering, aggregation, reporting, and visualization (Sharma et al., 2022). According to a recent systematic survey of existing LADs (Paulsen & Lindsay, 2024), 62% of the studied tools relied on descriptive analytics as their primary analysis method. Enabling instructors, administrators, or students with no background in data science to conduct descriptive analytics on their own has the potential to expand the reach of LA practice. However, to understand how GenAI can empower non-experts in conducting descriptive analytics within the domain of learning, it is essential to review the challenges non-experts face when engaging with this data analysis technique, the current way this analysis is conducted in LA practice, and the emerging role of AI-powered tools in facilitating data analysis. The following subsections address each of these topics in detail.

2.1 Barriers of Descriptive Analytics for Non-experts

Despite the potential of descriptive analytics to offer valuable insights, non-experts often face significant challenges in using these techniques effectively. The primary barrier is the complexity of data wrangling, which includes cleaning, organizing, and preparing raw data for analysis (Jaimovitch-López et al., 2023). While its importance could be disregarded by non-experts, data wrangling is the gateway to any analytic process (Furche et al., 2016) and its most time-consuming part (Singh et al., 2023). As such, learning the conceptual underpinnings and technical aspects of data wrangling is a fundamental part of any introductory course in data science (Gundlach & Ward, 2021). Most educators lack data-wrangling skills, which limits their ability to even start descriptive analytics processes (Mandinach & Gummer, 2016).

Even when the data has been pre-wrangled and is ready for analysis, the second major barrier is selecting the appropriate descriptive statistical methods to use. Non-experts often struggle to understand which techniques are best suited to summarizing their data, whether it be measures of central tendency (mean, median, mode), measures of variability (standard deviation, range), or frequency distributions (McGrath, 2014). These skills are typically acquired through college-level statistics courses and are not part of the standard training for most teachers or administrators.

Another high-level barrier is the ability to understand (let alone generate) medium- to high-complexity visualizations, which are often the outputs of descriptive analytics processes. While simple charts such as bar graphs or pie charts might be intuitive, more complex visualizations—such as box plots, scatter plots, or heatmaps—require a deeper understanding of data relationships and statistical concepts (Shreiner & Dykes, 2021). For non-experts, interpreting these advanced visualizations can be challenging, particularly when the data being visualized includes multiple variables, trends, or distributions that are not immediately apparent.

Finally, in addition to the conceptual knowledge needed for data wrangling, selecting appropriate descriptive statistics, and interpreting visualizations, non-experts must also have a certain level of proficiency in using software tools such as Excel, Tableau, or Power BI to perform these tasks (Carlisle, 2018). For individuals who lack familiarity with these platforms, the learning curve can be steep, further discouraging engagement with data analysis (Sá et al., 2024). The requirement to navigate software interfaces, understand tool-specific functions, and troubleshoot technical issues can be overwhelming for users who do not have a background in data science or analytics.

Any solution aimed at helping non-experts must address key challenges: simplifying data wrangling, guiding the selection of appropriate statistical methods, and enhancing the understanding of visualizations with contextual explanations. Additionally, it should be user-friendly, enabling stakeholders to interact with data easily through simple interfaces.

2.2 Descriptive LADs

Given the considerable barriers that non-experts have in conducting even descriptive analytics, the accepted solution in the field has been the LAD (Verbert et al., 2020). LADs usually take the form of a software application where different statistical analyses and visualizations are pre-determined—they are run online or offline over pre-wrangled data. The results are then presented to relevant stakeholders, particularly teachers, who are a natural audience for examining students' performance through analytics (van Leeuwen et al., 2022). By providing insights into students' learning processes and performance, LADs hold the potential to empower teachers to make informed decisions about their instructional practices. For instance, dashboards can enable teachers to monitor student engagement in online learning (Kaliisa & Dolonen, 2023; Dourado et al., 2021), support the orchestration of in-person class activities (Holstein et al., 2018; Martinez-Maldonado et al., 2012), and offer feedback to collaborative learning (Echeverria et al., 2024). These dashboards often include visual displays of students' activities and progress, which can inform instructional design and facilitate responsive feedback based on students' ongoing performance during class.

Despite this potential, researchers have found significant adoption challenges with teacher-facing dashboards. Teachers vary considerably in how frequently they use LADs and in how they respond to the information presented. To better understand these variations, researchers have proposed models to conceptualize teachers' use of dashboards, which often include phases such as awareness (van Leeuwen et al., 2021) and sense-making (Wise & Jung, 2019). These models reveal barriers to effective use, including teachers' limited data literacy (Pozdniakov et al., 2023) and differences in individual characteristics (van Leeuwen et al., 2021) or teaching experiences (Wise & Jung, 2019) that affect dashboard adoption. To address these issues, co-design approaches have been introduced, involving teachers directly in the design process to ensure that the dashboards prioritize contextual information that teachers find meaningful (Holstein et al., 2018; Martinez-Maldonado et al., 2022). While co-design has proven effective in making LADs more relevant and usable, it is a resource-intensive process, requiring researchers to work closely with teachers over extended periods to develop customized solutions (Sarmiento & Wise, 2022). Thus, while valuable, co-design is challenging to scale across diverse educational contexts.

A deeper underlying reason for these adoption challenges may be the primarily explanatory, rather than exploratory, design of most LADs (Echeverria et al., 2018). As mentioned in the previous section, data exploration through descriptive analytics is typically considered an expert-level practice. Explanatory LADs, on the other hand, aim to perform analytics tasks for teachers by presenting fixed visualizations and interpretations that simplify access to data. While this approach makes data more accessible, it also restricts teachers' agency to investigate and answer specific questions about their students. Moreover, explanatory LADs may contribute to an "over-fitting" issue, where teachers, limited by preset data representations, are more likely to confirm pre-existing biases or assumptions rather than uncover new insights (Campos et al., 2021). To truly empower teachers and promote data-informed instructional decisions, it is not just important but essential to incorporate more exploratory support within teacher-facing analytics tools, for example, using data storytelling techniques (Echeverria et al., 2018). However, building exploratory dashboards, given their requirement to be able to adapt to new questions and visualization types on demand and provide contextual explanations, is beyond the capabilities of traditional technologies. We propose that GenAI could serve as the technology that will enable us to build more exploratory dashboards that will allow non-expert teachers to conduct descriptive LA practice on their own.

2.3 Effectiveness of GenAI for Data Analysis

As mentioned in the introduction, several studies provide evidence that GenAI could be used to conduct data analysis tasks. While no studies focus directly on LA, in this subsection, we will discuss the findings of this recent body of knowledge from the adjacent fields of statistics and data science to better contextualize our analyses.

In the context of data science education, Zheng (2023) tasked students with several exercises where they needed to use ChatGPT v3.5 to answer different types of data science questions, from concepts and knowledge to coding and interpretation of results. For each exercise, students were asked to give their opinion about the usefulness of ChatGPT. The study found that ChatGPT performs well on simple problems but its performance diminishes with more complex and ambiguous questions. However, the performance can improve significantly with effective prompt engineering, highlighting the importance of crafting precise and informative prompts to guide the AI. The study also revealed that ChatGPT is helpful for understanding new concepts and clarifying existing knowledge, particularly when appropriate human prompts are used. Students reported positive experiences with coding assistance and explanations of coding parameters provided by ChatGPT, although the tool's effectiveness is heavily dependent on the user's ability to provide specific requests.

In the same context, Shen and colleagues (2024) evaluated ChatGPT's effectiveness in solving data science assignments from three different course levels. The study examined the use of raw prompts from course assignments and the application of prompt-engineering techniques to improve performance. They also found that ChatGPT performed well on simpler tasks but struggled with more complex ones. However, its success improved significantly with refined prompts, achieving high correctness rates.

In a study conducted on themselves, Owolabi and colleagues (2024) assessed the strengths and limitations of ChatGPT

in statistical data analysis. Their study used an econometric dataset with deliberate statistical issues to evaluate ChatGPT's performance. The results show that while ChatGPT can suggest appropriate statistical methods and provide guidance, it lacks the depth to replace a professional data analyst entirely. They recommend that the tool should be used by experts, but not by novices.

Connected with the main challenges of descriptive analytics for non-experts users described in Section 2.1, recent studies have focus on using LLMs to conduct several aspects of this process. For automating data-wrangling processes, Jaimovitch-López and colleagues (2023) demonstrated that LLMs like GPT-3 can significantly streamline data wrangling by efficiently transforming, standardizing, and extracting relevant data features. However, the study also identified challenges, particularly when dealing with tasks that require domain-specific knowledge or complex reasoning, such as unit conversions or semantic type detection. For visualization generation, Maddigan and Susnjak (2023) found that LLMs outperform existing natural language processing (NLP) techniques. However, their success is also highly dependent on the quality of the prompts provided. This underscores the critical role of user input in leveraging AI for effective data visualization. The study highlights that while AI tools can generate sophisticated visualizations, the user's ability to frame the right questions and interpret the results remains crucial.

The main insight drawn from these studies is that the usefulness of ChatGPT depends on the complexity of the problem, the user's level of expertise, and the quality of prompts used to guide it. However, none of the reviewed studies specifically or quantitatively assess the effectiveness of these AI tools for non-expert users, nor do they focus on descriptive LA. Given that the effectiveness of GenAI, particularly LLMs like ChatGPT, varies significantly across different domains and tasks (Lo, 2023), further research is needed to explore how these tools can be made accessible and beneficial for users with varying levels of expertise, especially in conducting descriptive LA. This gap is precisely what the present work aims to address.

3. Research Design and Methodology

In this section, we detail the approach taken to investigate the impact of previous technical expertise on the performance of individuals solving descriptive LA problems with the assistance of ChatGPT v4 and its Code Interpreter. Our study is guided by two primary research questions with respective hypotheses informed by related work:

- **RQ1:** Does prior level of expertise in data analysis affect the performance of individuals when solving descriptive LA problems of different difficulty levels with the support of ChatGPT v4 with Code Interpreter?
  - **Hypothesis:** The prior level of expertise in data analysis has a positive effect on performance, especially for harder tasks.
- **RQ2:** What other factors associated with previous experience and confidence, time on task, the characteristics of prompts used, and the actions conducted while solving the problems could explain the differences in scores obtained by individuals using ChatGPT v4 with Code Interpreter to solve descriptive LA problems?
  - **Hypothesis:** The most explanatory factors will be related to the prompts and the actions taken while solving the problems.

The first of these questions needs to be answered in an inferential manner; that is, it should provide statistically conclusive evidence of the effect (or lack thereof) of previous expertise at different difficulty levels, at least for medium-size effects. The second of these questions is exploratory in nature and should be answered using descriptive or inferential statistics or any other suitable data analysis method.

3.1 Experimental Design

To investigate the above research questions, we conducted a repeated-measures experimental study involving participants with varying levels of data analytics expertise. Each participant completed three tasks of varying difficulty using ChatGPT v4, with access to the Code Interpreter functionality. The Code Interpreter is a feature that enables ChatGPT v4 to execute Python code directly within the chat interface, extending its capabilities beyond text-based responses. This functionality allows users to perform a variety of data analysis tasks through natural language prompts, including data wrangling, summarization, aggregation, and visualization (Ahn, 2024). Users can upload datasets in formats like CSV or Excel and prompt ChatGPT in natural language, just as they would for any other ChatGPT function, to conduct data analysis.

The experiment was conducted from February to May 2024. Interested participants were required to take a screening test to determine their data analytics expertise levels, and they were then invited to join the study. After signing the institutional review board (IRB) consent form, participants were scheduled to participate in a 1.5-hour study conducted remotely via Zoom with an experimenter from the research team. The experiment process is illustrated in Figure 1. In this experiment, it was not feasible

to have a control condition, as novice participants would not be able to solve any of the tasks without the help of ChatGPT, while intermediate and expert participants could easily take 1 to 2 hours to solve the hardest task with traditional tools. Having a control condition with any tools other than an LLM would frustrate novices and require a large amount of effort from other participants.

In designing the experiment, we aimed to reduce factors that might potentially influence participants' performance. First, participants were reminded that the objective of the experiment was not to evaluate their performance but rather to assess ChatGPT's capabilities in assisting them with data analysis tasks. To encourage participants to make their fullest effort, they were informed that they could receive $25 for participating and an additional $25 bonus if they successfully solved all three questions. In reality, every participant was compensated $50 for their time. To discourage participants from copying and pasting prompts directly from the instruction document, we sent the instructions in image format. Considering that some participants might not have used ChatGPT before, we pre-recorded a tutorial video on using ChatGPT v4 for data analysis. The purpose of this video was to familiarize participants with the tool and reduce any potential anxiety. The video demonstrated how to upload data and provided simple examples of asking ChatGPT to summarize and visualize data. After watching the tutorial video and filling out a pre-survey on their demographics and prior experiences with ChatGPT, participants were given access to a paid ChatGPT v4 account.

Participants were then assigned three tasks of varying difficulty level: easy, medium, and hard, with time limits of 15 minutes, 20 minutes, and 20 minutes, respectively. During the experiment, participants were allowed to use any additional tools they preferred, in addition to ChatGPT. They were required to remotely share their screen with the experimenter, and the entire process was audio- and video-recorded. Participants were reminded to use a think-aloud protocol to help the experimenter better understand their thought processes and actions. Experimenters took detailed observation notes while participants worked on the tasks, but they refrained from intervening during the task. Upon completing the tasks, participants engaged in a 5- to 10-minute semi-structured interview about their experience using ChatGPT to solve analytical problems. After the experiment, participants were thanked and compensated for their time.

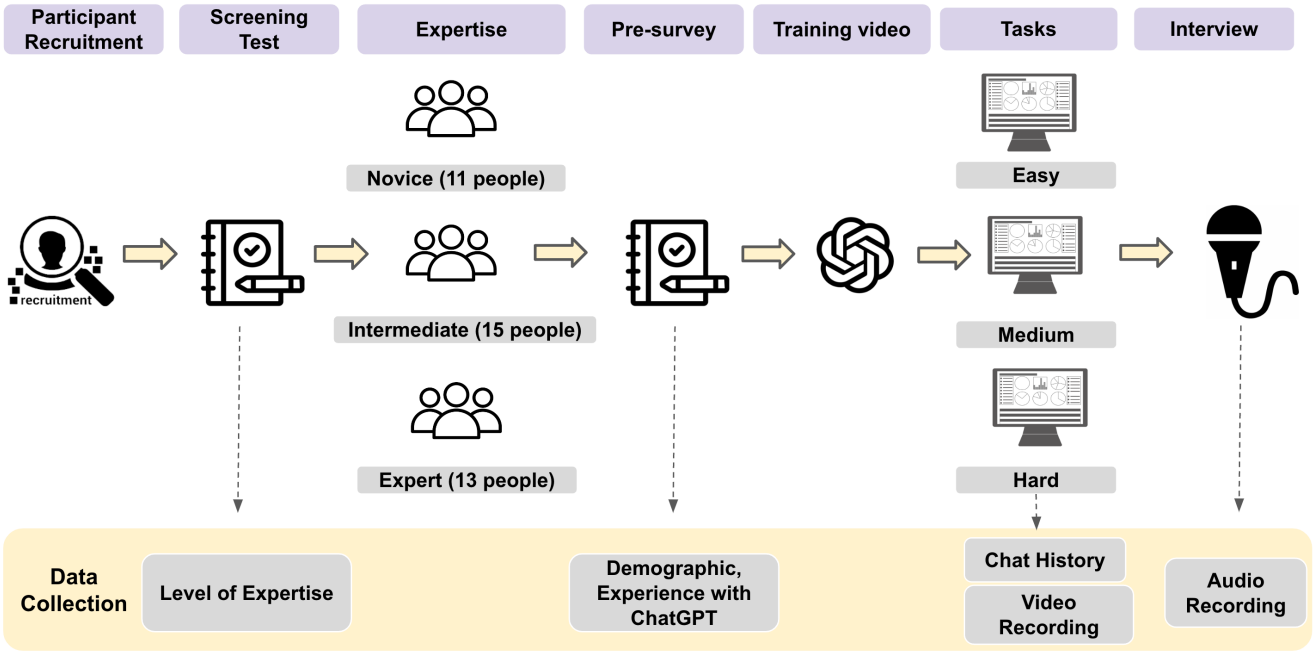

Figure 1. Experimental design and data collection.

3.2 LA Tasks

We summarized the three LA tasks in Table 1. All three tasks and the datasets used are available for review in the Open Science Framework (OSF) repository[1]. In designing these tasks, we had two considerations: educational context and analytics difficulty levels. We situated the LA in three educational contexts: from the perspective of a teacher who wants to identify struggling

[1] OSF repository: https://osf.io/4s9ng/?view_only=12b03b00d18d425380bc4a190e8d05d5

Table 1. Tasks and questions.

Task Summaries	Actions Needed per Question	Data Type	Observations
(Estimated Difficulty: Easy) Identifying struggling students and challenging subjects	1.1 Data filtering with one given condition	Continuous	40
	1.2 Data filtering with two given conditions		
	1.3 Calculate percentage based on previous filtering results		
	1.4 Evaluation of results in educational context		
	1.5 Calculate averages based on previous filtering results		
	1.6 Data visualization of the averages		
	1.7 Evaluation of results in educational context		
(Estimated Difficulty: Medium) Comparing collaborative learning performance	2.1 Data cleaning by changing data types and correcting typos	Continuous and categorical	40
	2.2 Data cleaning by removing empty values		
	2.3 Data classification on given conditions		
	2.4 Data visualization of percentages based on classified data		
	2.5 Evaluation of results in educational context		
	2.6 Calculate averages based on classification results		
	2.7 Explore relationship between a variable and the outcome		
(Estimated Difficulty: Hard) Analyzing online collaborative reading engagement	3.1 Generate network visualization from log file	Continuous, categorical, discrete, and time series	77
	3.2 Data filtering with self-defined conditions		
	3.3 Join two datasets, impute the empty values, and explore relationship between a variable and outcome	Continuous	12
	3.4 Join two datasets, and visualize group averages	Categorical	12

students and challenging subjects from students' test scores (easy task), from the perspective of a teacher who wants to test which learning method is more effective for their students (medium-difficulty task), and from the perspective of a professor who wants to analyze their students' online collaborative reading behaviours (hard task). We selected these situations as common examples in which descriptive LA tasks are typical. We also designed these tasks to vary in difficulty based on the complexity of the data processing and analytics involved in solving the questions. The difficulty levels of the tasks were not formally validated; however, each task reflects increasing levels of skill complexity as typically taught in regular data science courses (see Figure 1) and was used solely as conditions in a repeated-measures design rather than as an ordinal variable.

For the easy task, the questions are about data filtering based on one or two given conditions, calculating percentages and averages based on the filtered results, creating simple data visualizations, and evaluating the results in context. The medium-difficulty task involves data cleaning of different types of errors, removing empty values, classifying data, visualizing and calculating grouped data, and exploring the relationship between one variable and the outcome. The hard task involves generating a network visualization from a log file—a process that involves advanced data manipulation—defining conditions and then filtering the data based on the self-defined conditions, joining datasets, imputing empty values, and exploring the relationship between group variables and outcomes. The selection of different data analysis actions needed for different levels of difficulty was informed by traditional data science for educational data curricula (Estrellado et al., 2020) and the researchers' own experience as veteran instructors in LA courses.

3.3 Sample

Due to the requirement of providing a statistically conclusive answer to the first research question, the size of the sample was determined using statistical power calculation. The main statistical model to test the effect of previous experience on performance will be a repeated-measures ANOVA with within-subject (task difficulty) and between-subjects (level of expertise) variables. Both variables have three levels: task difficulty [easy, medium, hard] and level of expertise [novice, intermediate, expert]. The lower level of effect size that we wish to detect is medium ($f = 0.25$), with a traditional 0.05 significance level, a desired statistical power of $1 - \beta = 0.80$, and an expected correlation between measurements of 0.5. This level of statistical power required 36 participants.

The recruitment of participants was open to current school teachers, education professionals, and graduate students in education-related fields from surrounding communities and universities in a major US east coast city. All participants were recruited based on their experience in the educational field to represent the population of LA stakeholders, including teachers, instructional designers, and administrators. After completing the screening tests, 39 participants joined and completed the study. Among these participants, 27 identified as female and 11 as male, with ages ranging from 22 to 44 and a median age of 25. All participants held a bachelor's degree and were currently graduate students in an educational technology program. Three of the participants also worked as teachers in local schools. All participants were fluent in English, and none were currently enrolled in an LA course.

3.4 Data Sources and Preprocessing

We administered an online screening test and a pre-survey prior to the experiment, both of which are available for review in the OSF repository. We saved and transcribed the chat histories generated by 39 participants, resulting in a total of 973 prompts. Additionally, we video-recorded the entire study. The total recordings exceed 60 hours, from which a total of 2175 participants' actions were identified. We summarized the data collected in Table 2. Specifically, participants' levels of expertise were assessed prior to the study (see Section 3.4.1). Their demographics, previous experience using ChatGPT, and self-perceived confidence in using ChatGPT for analytics tasks were collected in the pre-survey. Participants' performance on the three tasks was graded by two trained graders who reviewed video recordings and chat histories according to a scoring rubric (see Section 3.4.2). The time each participant spent on each task was transcribed from video recordings. The prompts were transcribed from chat histories, and we counted the total number and length of prompts. Three trained coders qualitatively assessed the originality and sophistication levels of each prompt (see Section 3.4.3). Additionally, two members of the research team reviewed all video recordings and qualitatively coded all actions taken by participants, as well as their thought processes derived from the think-aloud data during the study (see Section 3.4.4).

Table 2. Data source.

Category	Data	Source
Prior factors	Level of data analytics expertise	Screening test
	Previous experience with ChatGPT	Pre-survey
	Confidence of using ChatGPT for analytics	Pre-survey
Performance	Scores achieved	Chat history and video
	Time on task	Video
Prompts	Total number of prompts	Chat history
	Length of each prompt	Chat history
	Originality level	Chat history
	Sophistication level	Chat history
Process	Actions being taken	Video
	Thinking process	Video

3.4.1 Screening Test

An experienced instructor in LA (with 10 years of teaching experience in the field) designed a screening test to assess the participants' data analysis expertise. The test comprised six open-ended questions. The first two questions were designed to evaluate whether participants could filter data and select simple visualizations—skills necessary for solving the easy task during the experiment (see Table 1). The first question asked participants to identify the best way to visualize the difference in final scores between two groups of students. The second question, using a simple dataset containing information on student names, ages, grades, enrolments, and extracurricular activities, required participants to explain how to generate a sublist of students who were at least 15 years old, in grades 10 to 12, and part of the chess club using Excel or another software tool.

The next two questions tested participants' abilities in data wrangling and data aggregation—skills corresponding to the ones needed for the medium-difficulty task. The first question asked how to handle a dataset with 10% missing data in a column. The second question presented a dataset with columns for ID number, name, class, subject, and test scores and asked participants how to calculate the average score for each class and subject.

The final two questions assessed participants' skills in joining data from multiple datasets and exploring relationships between variables—abilities required for the hard task. The first question involved merging two datasets: one containing student IDs, names, subjects, and test scores, and another with student IDs, names, total classes, and classes attended. Participants were asked to explain how to join these datasets. The second question required participants to use the merged dataset to determine if there was a relationship between class attendance and final scores.

The instructor that designed the test reviewed participants' responses to assess whether they demonstrated conceptual understanding and the ability to solve each task. The responses were classified as follows: participants who showed proficiency in the first and second groups of questions and in at least one question from the third group were classified as experts; those proficient in the first group and at least one question from the second group (but not the third) were classified as intermediate; the rest were classified as novices. If a participant struggled with an easier question but succeeded in more challenging ones, the instructor made a final judgment. This special review was needed for three participants, resulting in two being reclassified as experts and one as intermediate.

During recruitment, it was made clear to participants that their eligibility for the study was based not on their answers but on the availability of spots. Out of all registered volunteers, we invited 53 participants to the experiment (19 experts, 18 intermediates, and 16 novices). Thirty-nine participants (13 experts, 15 intermediates, and 11 novices) agreed to participate in the study.

3.4.2 Performance Grading

We developed a grading rubric for each question on a 1 to 5 scale. The general principle was to grade the result first, whether it came from ChatGPT output or other tools the participants chose to use. If the result was correct, we awarded 5 points. If the result was incorrect, we graded based on the partial correctness of their answers, as reflected by the participants' prompts or their thinking process if they did not use ChatGPT. If there were minor errors in the prompts, we awarded 3 or 4 points; if there were conceptual errors, we awarded 2 points; and if the prompt was completely incorrect, we awarded 1 point. If a participant gave up on answering the question, we awarded 0 points.

The grading process involved three steps. First, we trained the graders. Two graders were given the first two participants' data and graded the questions using the rubric. Additionally, two expert graders also graded the same data. We then met to discuss the differences in grades. After reaching a consensus, the two graders independently assessed 23% of the data. Since the grades are ordinal, we calculated the weighted Cohen's Kappa (Cohen, 1968), which is typically used to measure agreement between two raters for ordinal data. The weighted Cohen's Kappa was 0.74, indicating substantial agreement (McHugh, 2012). Finally, the two graders evenly split the remaining data and graded it independently.

3.4.3 Prompt Coding

To better understand how participants developed their prompts and assess the extent to which they used their additional statistics or data science background knowledge, we applied an iterative process to develop two coding schemes: one for classifying the prompt originality (see Table 3) and the other for evaluating the prompt sophistication level (see Table 4). The coding process involved three coders who independently applied the coding schemes to all data. To ensure a common understanding and refine the coding schemes with specific examples, we conducted three rounds of coding discussion meetings. Before each meeting, coders independently coded approximately 5% of the data and brought their questions and observations to the meetings. These discussions helped refine and finalize the coding schemes, and the final codes were selected upon reaching consensus. After finalizing the coding schemes, to measure the inter-rater reliability, the three coders independently coded an additional 12% of the data.

Table 3. Prompt originality.

Method	Definition	Examples
Copy	Prompts were exactly the same as the given questions.	"List all the students that 'struggled' in at least one course in the previous year."
Modification	Prompts were based on given questions but rephrased in participants' own words.	"For the columns of 'quiz 1 score' and 'quiz 2 score,' convert text into corresponding integer, for example, convert 'eighty' to '80.'"
Original	Participants created their own prompts.	"What is network visualization?" "Can you make it less cluttered?"

Table 4. Prompt sophistication.

Levels	Definition	Examples
Low	Prompts used everyday language or were copied from the given questions.	"Turn the scores into number format if it's a word."
Medium	Prompts included one statistical or data science term not in the original questions.	"Firstly in Quiz 1 Score and Quiz 2 Score, replace all non numeric values with the equivalent numeric value."
High	Prompts employed statistical or data science concepts and terminology, demonstrating a deep understanding of these concepts.	"Create a network visualization using the processed data. The nodes are Name, the edges are identical data in Range. If two Name has the same Range, there is an edge."

Considering that we had three coders and used nominal codes for prompt originality and ordinal codes for sophistication level, we measured inter-rater reliability by calculating Fleiss's Kappa (Fleiss, 1971) for prompt originality and Kendall's coefficient of concordance (Field, 2005) for sophistication level. Fleiss's Kappa value was 0.522 for prompt originality, indicating fair agreement among coders. For sophistication level, Kendall's coefficient of concordance was 0.531, indicating moderate agreement. Recognizing that these agreement levels were not substantial, we held a fourth meeting to address discrepancies and reach a consensus on final codes. To ensure high-quality coding, all remaining prompts were coded by all three coders. The final code for each prompt was determined by calculating the mode of the three codes. Agreement was achieved by at least two coders for 89% of originality and 96% of sophistication. For prompts where no mode was available, discrepancies were resolved by an expert coder.

3.4.4 Process Coding

To better understand participants' processes of conducting data analysis, we adapted the coding scheme for exploratory data analysis developed by Daele and Janssenswillen (2022). The finalized coding scheme consists of six major processes: understanding the problem, data wrangling, data exploration, data analysis, meta-cognitive steps, and other processes. Within each category, two to four specific actions were coded from participants' prompts, operational actions, and think-aloud data from video recordings. We detail this coding scheme in Table 5.

The coding process began with two coders independently coding the first 20% of the data with the unit of analysis defined as individual actions—specific steps taken by participants during problem-solving. These actions include discrete steps such as prompting ChatGPT; interacting with files and other programs; and verbal reasoning steps like articulating next steps, evaluating results, or reflecting on findings. For example, when participant 7 was working on a question in task 1, they initially prompted ChatGPT to "give me the % of students that struggled in math." After viewing the output, they opened the dataset in Excel to compare the result and then said, "I don't think 10% is a significant number," as an evaluation of the result. In this case, three actions were coded in sequence: "summarize data," "check results," and finally "evaluate results." To ensure consistency, both coders documented detailed descriptive notes on each action to facilitate comparison and the consensus-building process. To assess the reliability of multiple coders in annotating sequences of actions, we calculated the inter-rater reliability agreement (IRRA) as proposed by Sullivan (2014). IRRA was developed to measure inter-rater reliability in evaluating students' pathways in problem-solving and is based on the minimum edit distance (Levenshtein distance), which quantifies the number of insertions, deletions, and substitutions required to transform one sequence into another. We adopted this approach because it is more suitable for processing sequential data in string format than traditional kappa metrics that measure categorical variables (Sullivan, 2014). The IRRA ranges from 0 to 1, with 1 indicating perfect agreement. Once we reached an IRRA of 0.735, indicating a high level of agreement between the two coders (Sullivan, 2014), based on the 20% of the data, the two coders split the remaining data and graded it independently.

4. Analysis and Results

The analysis conducted in this work is divided into two parts, corresponding to our research questions. In the first part, we perform analyses to determine if there are statistically significant differences in the scores obtained in the descriptive LA tasks among participants with different levels of data analysis expertise. In the second part, we use statistical analysis and process visualizations to explore which factors explain the differences (or lack thereof) in the scores of the different participants.

4.1 Relationship between Expertise and Score

Before we start answering the research questions, we descriptively analyze the scores obtained by participants to provide context for the following analyses. Each question was graded from 0 to 5 according to the method described in Section 3.4.2.

Table 5. Processes and actions coding scheme.

Processes	Actions	Definition	Examples
1. Understanding Problem	1.1 Read assignment.	Participant opened the task and read it.	N/A
	1.2 Explore original data.	Participant viewed the data file with Excel, or prompted to summarize the original data.	Participant prompted "show data as a table."
2. Data Wrangling	2.1 Load original data.	Participant uploaded the data file to ChatGPT or other tools.	N/A
	2.2 Data cleaning	Participant cleaned the data.	Participant prompted to convert non-numeric data to numeric data.
	2.3 Data manipulation	Participant manipulated the data (e.g., grouping, filtering, joining, etc.).	Participant prompted to join three datasets.
	2.4 Load manipulated data.	Participant uploaded or downloaded the cleaned dataset to or from ChatGPT or other tools.	N/A
3. Data Exploration	3.1 Summarize data.	Participant performed descriptive statistics.	Participant prompted "how many students struggled in math?", "what is the average?"
	3.2 Visualize data.	Participant visualized data or fine-tuned visualization.	Participant prompted "can you give me a graph?"
	3.3 Other exploratory analysis	Participant ran regression, correlation, or other testing.	Participant prompted to run a multivariate regression analysis.
4. Data Analysis	4.1 Evaluate results.	Participant made sense of results in the education context and reached a conclusion.	Participant said "I don't think it is a significant number because. . . ."
	4.2 Reasoning steps	Participant thought about the steps to solve the problem.	Participant said "I want to know the total number of students and calculate the percentage."
5. Meta-cognitive Steps	5.1 Consult additional information.	Participant sought additional information to help them understand and solve the questions.	Participant prompted "what is network visualization?"
	5.2 Check results.	Participant checked if ChatGPT was correct.	Participant checked the answer by eyeballing the spreadsheet, recalculated with calculator or other software.
6. Other	6.1 Upload a screenshot.	Participant uploaded the screenshot of the instructions.	N/A

The scores obtained in each question were added and normalized to obtain a score from 0 to 10 in each task. The total score of the test was calculated by summing the scores of the three tasks; therefore, the theoretical range of the total score was from 0 to 30 points. The distribution of values for the total score and the score per task can be seen in Figure 2. The average total score was 22.35 [$SD = 5.34$]. The distribution of scores for the different tasks was as follows: easy [$M = 8.19, SD = 2.11$], medium [$M = 7.19, SD = 1.89$], and hard [$M = 6.96, SD = 3.14$]. To find answers to RQ1, we used two additional sources of data: (1) the level of expertise assigned to each participant after the screening test and (2) the level of difficulty of each task.

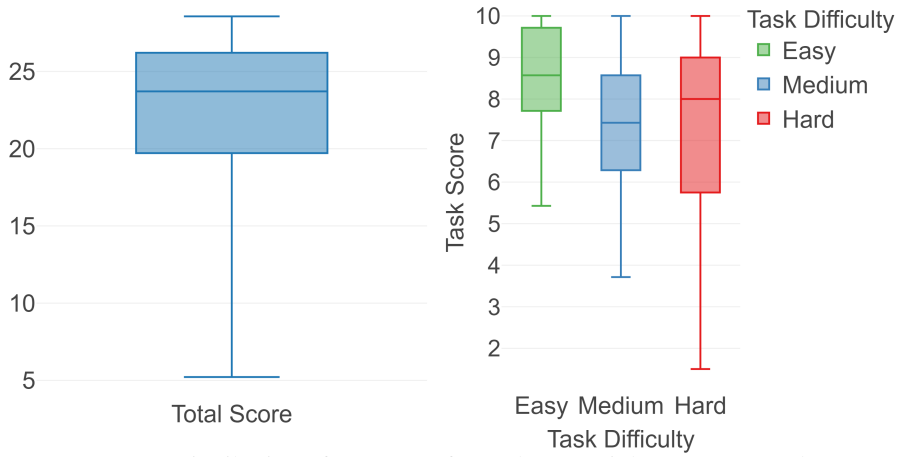

Figure 2. Distribution of scores. Left: total score; right: score per task.

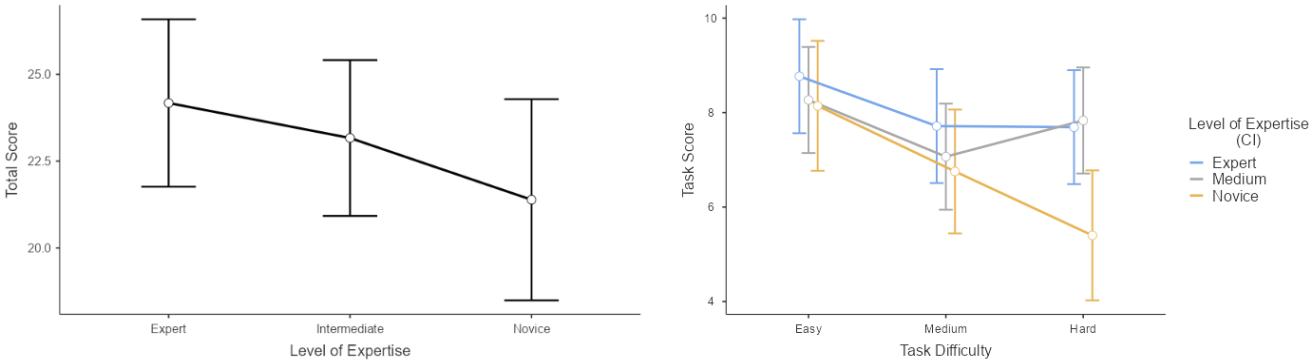

Figure 3. Left: Total score obtained by participants with different levels of expertise. Right: Task score at different task difficulty levels; each line represents a different level of expertise. In both graphs, error bars represent 95% confidence intervals.

4.1.1 Total Score

This analysis used the total score obtained in the whole test. Two participants (both novices) were excluded from the analysis due to missing data for at least one task. Unsurprisingly, expert participants achieved the highest average score [$M = 24.18, SD = 4.60$], followed by intermediates [$M = 23.17, SD = 3.61$], and finally novices [$M = 21.40, SD = 4.81$]. It is interesting to note that the difference between the average scores of experts and novices was less than 3 points, or less than 10% of the test range (see left side of Figure 3). To determine if these differences in scores were statistically significant, we used a one-way ANOVA test. After verifying that the data and residual distributions met the test's assumptions, we analyzed the results. The test did not find a significant effect of the independent variable (IV), level of expertise, on the dependent variable (DV), total score [$F(2,34) = 1.13, p = 0.334$]. The achieved statistical power of the test was very high [$1 - \beta = 0.98$]. These results are a preliminary indication that, when using ChatGPT and Code Interpreter as support tools, participants with different levels of expertise could achieve similar levels of performance in the execution of descriptive LA tasks.

4.1.2 Score per Task

To determine whether task difficulty had any effect on performance across different expertise levels, this analysis was conducted using the score of each task individually. A visual inspection of the distribution of the scores (see right side of Figure 3) shows that the performance pattern observed in the analysis of the total grade applies only to easy and medium-difficulty tasks, where the difference in mean task score is less than 1 point (10% of the grade) between participants of different expertise levels. However, in the hard task, the average score of novices [$M = 5.4, SD = 3.40$] is considerably lower than that of intermediate [$M = 9.0, SD = 2.06$] and expert [$M = 9.0, SD = 3.15$] participants.

To test whether the differences observed across varying levels of task difficulty and participant expertise were statistically significant, we ran and tested a linear mixed-effect model (LMM). We used this model instead of a repeated-measures ANOVA because our data violated the sphericity requirement of ANOVA (Mauchly's W test was significant with $p = 0.005$) and because there were two missing values for specific tasks. LMMs are more lenient toward non-sphericity and do not discard participants if one measurement is missing, unlike ANOVA. Otherwise, LMMs are functionally equivalent to a repeated-measures ANOVA

and provide more statistical power in our specific situation (Wallace & Green, 2013). In this LMM we use the participant identification as the cluster variable responsible for the random effect in the intercept of the linear function. The analysis found that both independent variables, task difficulty [$F(2,72) = 4.34, p = 0.010$] and level of expertise [$F(2,36) = 4.89, p = 0.020$], had a statistically significant effect on the differences in task scores. However, the test did not find a significant interaction between the independent variables [$F(4,72) = 0.957, p = 0.437$]. Post hoc Bonferroni-corrected pairwise tests for task difficulty levels found significant differences in the scores obtained by all participants only between the easy and difficult tasks [$p = 0.019, g = 0.536$]. The same tests for level of expertise found significant differences only between experts and novices [$p = 0.026$]. The achieved statistical power of the test was high [$1 - \beta = 0.84$] and in line with our original design. While the result is statistically significant, the difference between the scores of experts [$M = 8.1, SD = 2.1$] and novices [$M = 6.8, SD = 2.7$] remains small (less than 1.5 points out of 10 per task on average). This analysis also shows that the scores on the hard task [$M = 7.1, SD = 3.0$] are not significantly lower than those on the medium-difficulty task [$M = 7.2, SD = 1.9$] (see Figure 3).

However, the most interesting finding of this analysis is the very low fit of the LMM. Both task difficulty and level of expertise account for only 16% of the marginal variability in the scores [Adj. $R^2 = 0.164, LTR\chi^2(8) = 21.230, p = 0.007$]. This result also supports the idea that the successful use of ChatGPT does not strongly depend on the user's previous level of expertise, opening the door for facilitating LA practice to individuals with educational domain knowledge but lacking data science skills.

4.2 Performance Clustering

Given that the level of expertise of the participants does not fully explain the variability in the scores, the second part of the analysis will explore other factors related to the participants and their use of ChatGPT that could better explain their different levels of performance. The first step in this exploration is to create an explainable grouping of participants according to their performance. In this analysis, we used only the data from the scores obtained for each task per participant.

A three-dimensional vector space was created using the participants' scores, with each dimension corresponding to the level of difficulty of the question (easy, medium, hard). A K-means clustering algorithm was applied to this space to produce three performance clusters. The number of groups, three, was selected based on the recommendation from the elbow method (the change in slope occurred between three and four groups), the interpretability of the results (when four clusters were selected, there were two very similar groups), and the desire to avoid having clusters with too few participants (when four clusters were selected, one cluster had fewer than five members). The box plot of the scores of their members can be seen on the left side of Figure 4. The interpretation given to the clusters was as follows:

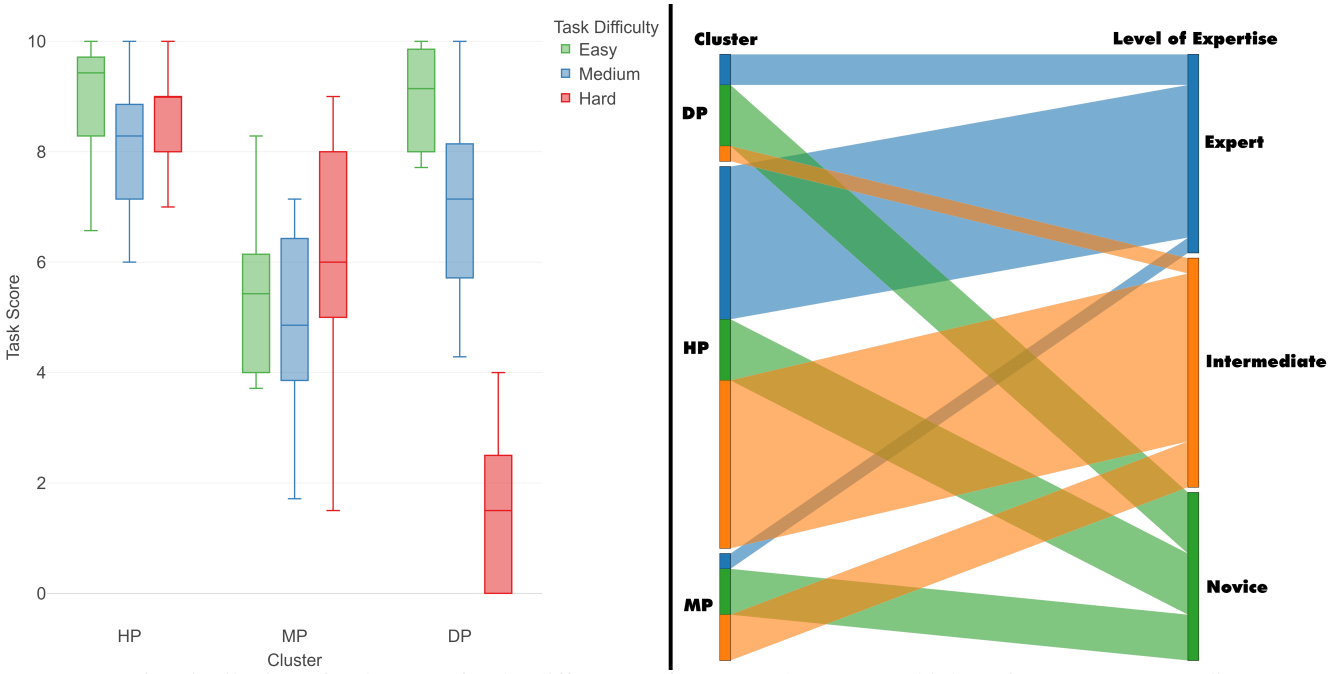

Figure 4. Left: Distribution of task scores for the different performance clusters. HP: high performance, MP: medium performance, DP: decreasing performance. Right: Cluster composition based on levels of expertise original classification.

- **High performance** (HP; 25 participants): Obtained average scores of 9 in all tasks.

- **Medium performance** (MP; 7 participants): Obtained average scores of 5 for the easy and medium tasks, improving to 6 in the hard task.
- **Decreasing performance** (DP; 7 participants): Obtained average scores of 9 for the easy task, 7 for the medium-difficulty task, and 2 for the hard task.

The composition of the clusters based on the level of expertise of the participants can be seen on the right side of Figure 4. It is interesting to note that most experts and intermediate participants are part of the HP cluster. Only three out of 13 experts and four out of 15 intermediates are not in HP. However, the 11 novices are almost evenly distributed among the groups (four in HP, four in DP, and three in MP). This diversity in the performance groups of novices is another indication that the score obtained is not necessarily determined by the previous level of expertise. This reinforces the hypothesis that the difference lies in how the participants used ChatGPT, for example, the types of prompts they used or the sequence of actions taken while solving each task. To confirm that this grouping better explains the differences in performance, the same LMM was used, replacing the IV factor level of expertise with the newly obtained performance cluster. As expected, different performance clusters show large statistically significant differences in scores among themselves [$F(2,36) = 37.7, p < 0.001$] (see Figure 5). However, the LMM shows that there is an interpretable way to group the participants that explains the variance in the results better than previous expertise. What we do not know are the characteristics of the participants that belong to these clusters apart from their performance. The following analyses will explore how other participants' variables could be used to explain the scores or the membership to different performance clusters.

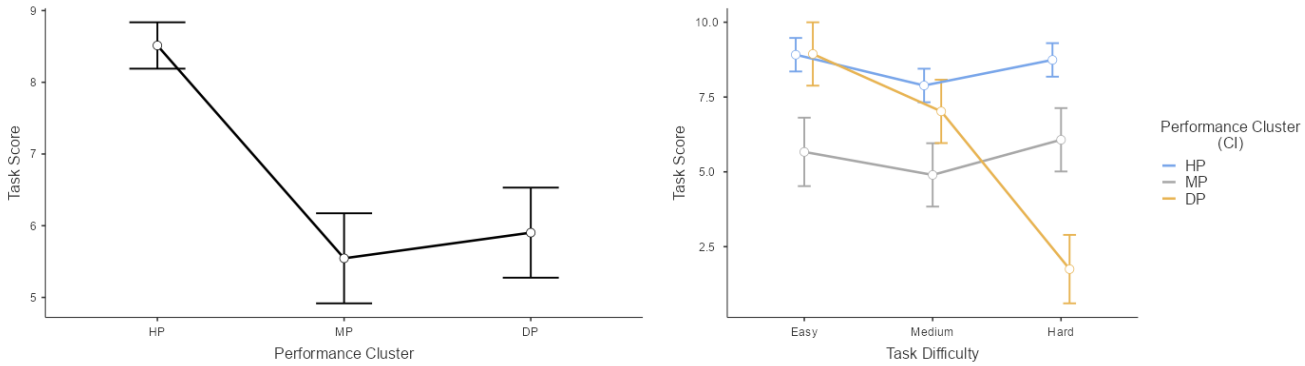

Figure 5. Left: Total score obtained by participants in different performance clusters. Right: Task score at different task difficulty levels; each line represents a different performance cluster. In both graphs, error bars represent 95% confidence intervals.

4.3 Previous Experience and Confidence

During the pre-survey, participants were asked about their perception of (1) their proficiency with ChatGPT for solving general tasks [1–5 Likert scale], (2) the frequency with which they use ChatGPT [1–5 Likert scale], (3) if they have used ChatGPT to solve data analysis problems [True/False], and (4) their level of confidence that they will be able to successfully solve data analysis problems with ChatGPT [1–5 Likert scale]. The distribution of these variables, with indications of the original expertise classification, can be seen in Figure 6. This analysis used this data to attempt to predict participants' task scores or explain their membership in the different performance clusters.

4.3.1 Predicting Score

Employing the same LMM that was used to determine the effect of level of expertise and task difficulty on the task scores, we replaced each experience and confidence variable. As a result of using Likert scales, all of these variables were added as factors, not covariates. The results were not significant for all the variables: proficiency [$F(4,33.8) = 0.720, p = 0.585$], frequency of use [$F(4,33.8) = 0.276, p = 0.891$], use for analytics [$F(1,34) = 0.711, p = 0.405$], and confidence [$F(4,31.0) = 0.254, p = 0.905$]. Also, their interactions with task difficulty were non-significant. All of these variables were introduced in the model as ordinal independent factors.

Given that by themselves these variables did not significantly explain the variance in scores, we tested, using a nested LMM, if adding them to the original level of expertise and difficulty level model used in Section 4.1.2 helped to explain more of the task score variance. The result, again, was non-significant increments in the Adjusted R^2 of the model for all variables. These two results combined lead to the conclusion that the previous experience and confidence variables either overlap significantly with expertise or do not provide information that can explain the variability in the scores.

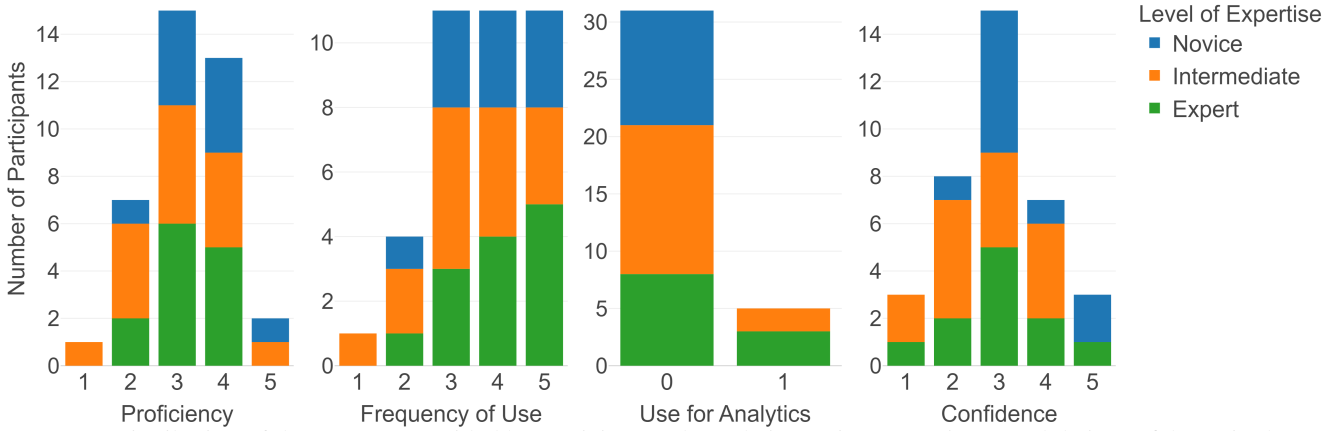

Figure 6. Distribution of the answers provided by participants about their previous experience and their confidence in the use of ChatGPT.

4.3.2 Predicting Performance Cluster Membership

To test if the previous experience and confidence variables can predict the participant's membership in one of the performance clusters, we used a multinomial logistic regression. In this case, performance cluster was the DV and the experience and confidence variables were the IVs. Because we wanted to know what differentiates medium and declining performers from high performers, the HP cluster was used as a reference. We found that the different levels of these variables also do not significantly differentiate between performance clusters: proficiency model $[\chi^2(8, N = 38), R^2_N = 0.238, p = 0.109]$, frequency of use model $[\chi^2(8, N = 38), R^2_N = 0.0946, p = 0.755]$, use for analytics model $[\chi^2(2, N = 36), R^2_N = 0.001, p = 0.975]$, and confidence model $[\chi^2(8, N = 36), R^2_N = 0.184, p = 0.153]$. A combined model, including the interactions between proficiency and frequency of use and proficiency and use for analytics, was also non-significant $[\chi^2(38, N = 36), R^2_N = 0.776, p = 0.158]$. This result, combined with those obtained in the previous subsection (Section 4.3.1) and those made for level of expertise, strongly suggests that previous experience either with data analysis or with ChatGPT is not a good predictor of the performance levels that participants can achieve.

4.4 Time on Task and Prompt Characteristics

Given that previous experience and confidence levels before the experiment do not explain the scores, we will now focus on the interactions that occurred during the experiment. First, we will consider the time on task [continuous] that the participants took to complete each task and the calculated and coded characteristics of the prompts used per task: number of prompts [continuous], total length [continuous], originality [copy, modification, original], and sophistication [low, medium, high]. The distribution density of time on task, disaggregated by task difficulty and performance cluster, can be seen in Figure 7. The distribution density of the number of prompts and total length separated by performance cluster can be seen in Figure 8. The percentage of use of each level of originality and sophistication per performance cluster can be seen in Figure 9. Given that both originality and sophistication could have different levels for the prompts of each task, an originality score and sophistication score were obtained by a weighted sum of their frequencies. Additionally, the average length of a prompt (average length) was calculated by dividing the total length by the number of prompts to use in individual models.

4.4.1 Predicting Score

First, we tested if the time on task or the characteristics of the prompts used by the participants had an influence on the scores obtained. We again used an LMM factored by task cifficulty. We found that time on task (covariant) $[F(1, 101.5) = 0.129, p = 0.720]$, number of prompts (covariant) $[F(1, 103.9) = 3.65, p = 0.059]$, average length (covariant) $[F(1, 100.8) = 2.37, p = 0.127]$, originality score (covariant) $[F(1, 99.4) = 0.898, p = 0.346]$, and sophistication score (covariant) $[F(1, 70.9) = 0.245, p = 0.622]$ were not statistically significant predictors of the variance in the score. Also, none of these variables interact significantly with task difficulty. On the other hand, we found a statistically significant effect of total length (covariant) $[F(1, 101.3) = 7.38, p = 0.008]$. No significant interaction was found between total length and task difficulty.

An analysis of the coefficients of the model shows that an increase of one additional word increases the predicted score by 0.007 points. Given that the standard deviation of total length is 85 words, that could correspond to a change of 0.7 points out of 10 in the task grade, or one-third of the observed standard deviation in the same scores. This result seems to indicate that the total amount of words used to solve a problem is connected with the score obtained by the participants.

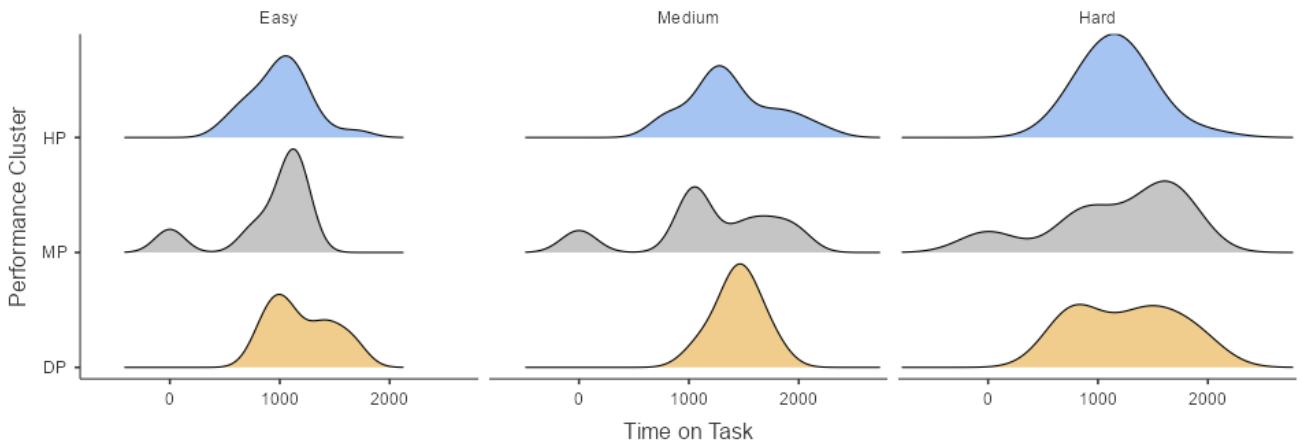

Figure 7. Distribution density of time on task disaggregated by task difficulty and performance cluster.

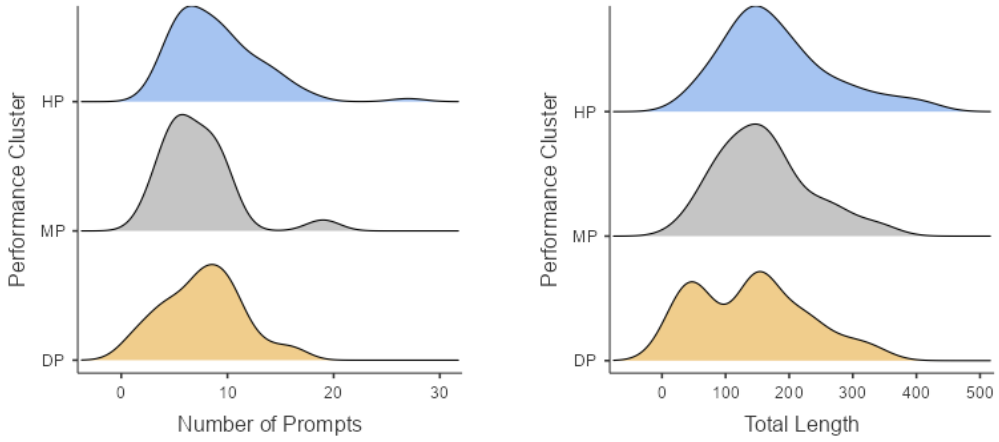

Figure 8. Left: Distribution density of number of prompts. Right: Distribution density of total length of prompts. Both densities are grouped by performance cluster.

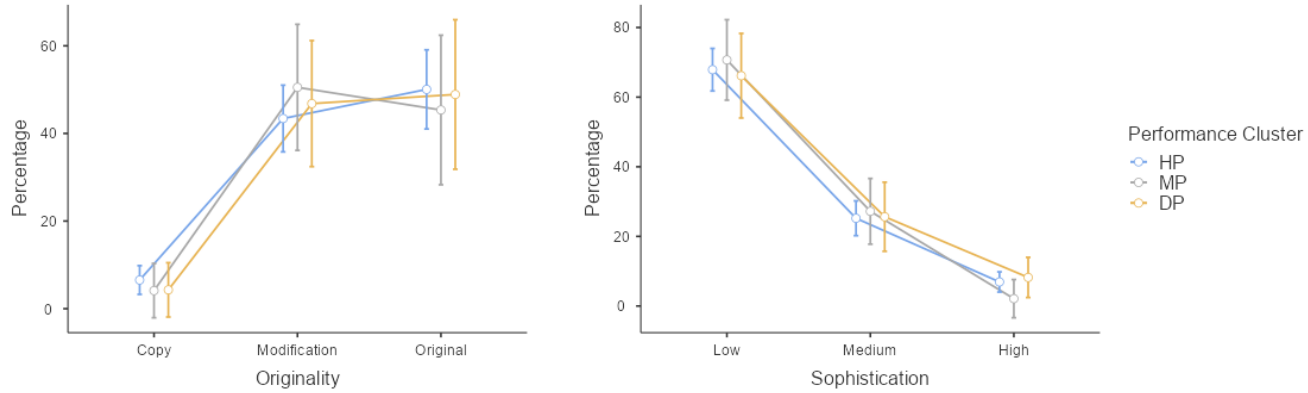

Figure 9. Left: Percentage of use of each level of originality. Right: Percentage of use of each level of sophistication. Both graphs are disaggregated by performance cluster.

4.4.2 Predicting Performance Cluster Membership

A multinomial logistic regression was used to determine if the time on task or the characteristics of the prompts used by the participants could predict which performance cluster they belong to. The variable average length was excluded from this model as it is a linear combination of two other variables. Variables connected with effort were statistically significant predictors of

membership in a given performance cluster: time on task [$\chi^2(2, N = 115) = 7.26, p = 0.027$], number of prompts [$\chi^2(2, N = 115) = 7.08, p = 0.029$], and total length [$\chi^2(2, N = 115) = 10.46, p = 0.005$]. Variables related to the language used in the prompt did not significantly predict the membership: originality score [$\chi^2(2, N = 115) = 1.70, p = 0.427$] and sophistication score [$\chi^2(2, N = 115) = 5.10, p = 0.078$]. An analysis of the coefficients shows that the variable number of prompts helps to statistically explain the difference between high performers and medium performers [odds ratio $= 0.71, p = 0.015$]. That means that for each additional prompt per task, the probability of being in HP instead of MP increases by 40% on average. It also shows that the variable total length helps to statistically explain the difference between high performers and declining performers [odds ratio $= 0.984, p = 0.023$]. That means that for every additional word used during the task, the probability of being in HP instead of DP increases by 1.6% on average. The complete visual representation of the probabilities versus the value of these variables can be seen in Figure 10. These results suggest that the difference between performance clusters is in part explained by the number and length of the prompts used, more than the actual originality or sophistication of the language.

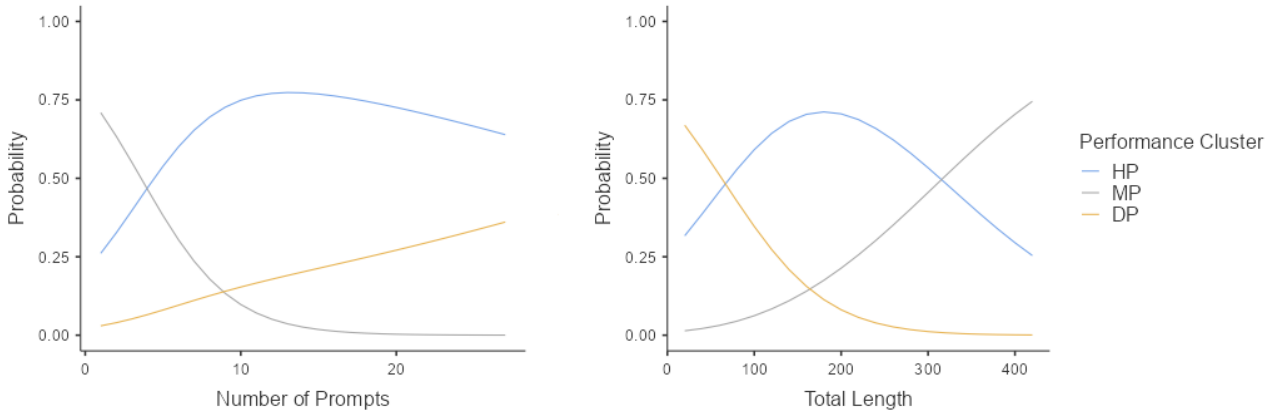

Figure 10. Probability of performance cluster membership given the values of the variables. Left: number of prompts; right: total length.

4.5 Action Sequence and Process Analysis

The richer source of data obtained for this work is the sequence of actions performed by the participants while solving the tasks. Due to the inadequacy of traditional statistical models to deal with sequences of variable length, we applied sequence analysis, process mining, and network analysis tools to explore how these sequences could help us interpret the differences in task score and performance cluster membership. We started with visualizing the raw sequences to gain insights into their composition. These visual representations, grouped by action category and separated by performance cluster for the easy task, can be seen in Figure 11. The analysis of the difference between different performance clusters in these visualizations first pointed to longer sequences in the HP cluster than in the other two clusters, especially in DP. Also, sequences in the HP cluster seem to be "busier"; that is, they have more transitions over different types of actions than those in the other two clusters. No other insight was gained from this visual inspection.

To better understand the transitions between actions, a heuristic mining algorithm was used to construct a process map for actions taken by each performance cluster and during each task. The high-level overview of the process map for the hard task is shown in Figure 12. This high-level view illustrates again the difference in how complex these maps are, reinforcing the interpretation of the visual analysis of the sequences. The process maps also provide descriptive information on the percentage of participants in each cluster that performed an action or transitioned from one specific action to another, which could also lead to potential differences between clusters. In the following subsections, we investigate both the complexity of the sequences and the prevalence of actions and transitions more formally.

4.5.1 Sequence Length and Complexity

The length of the sequence can be easily calculated and is just the number of different steps taken by the participants as determined by the coders. To measure how "busy" or "complex" a sequence is, we will use the complexity index developed by Ritschard (2023). This index counts and normalizes the number of transitions in a sequence and combines it with a normalized longitudinal entropy (logarithm of the inverse of different states) to estimate the complexity produced by the state distribution in the sequence. We calculated the complexity index for each sequence produced by the participants for each task. The complexity index varies from 0 (just one step in the sequence) to 1 (no two sequential states are the same, and all possible states are visited equally). The mean and standard deviation for these values can be seen in Table 6.

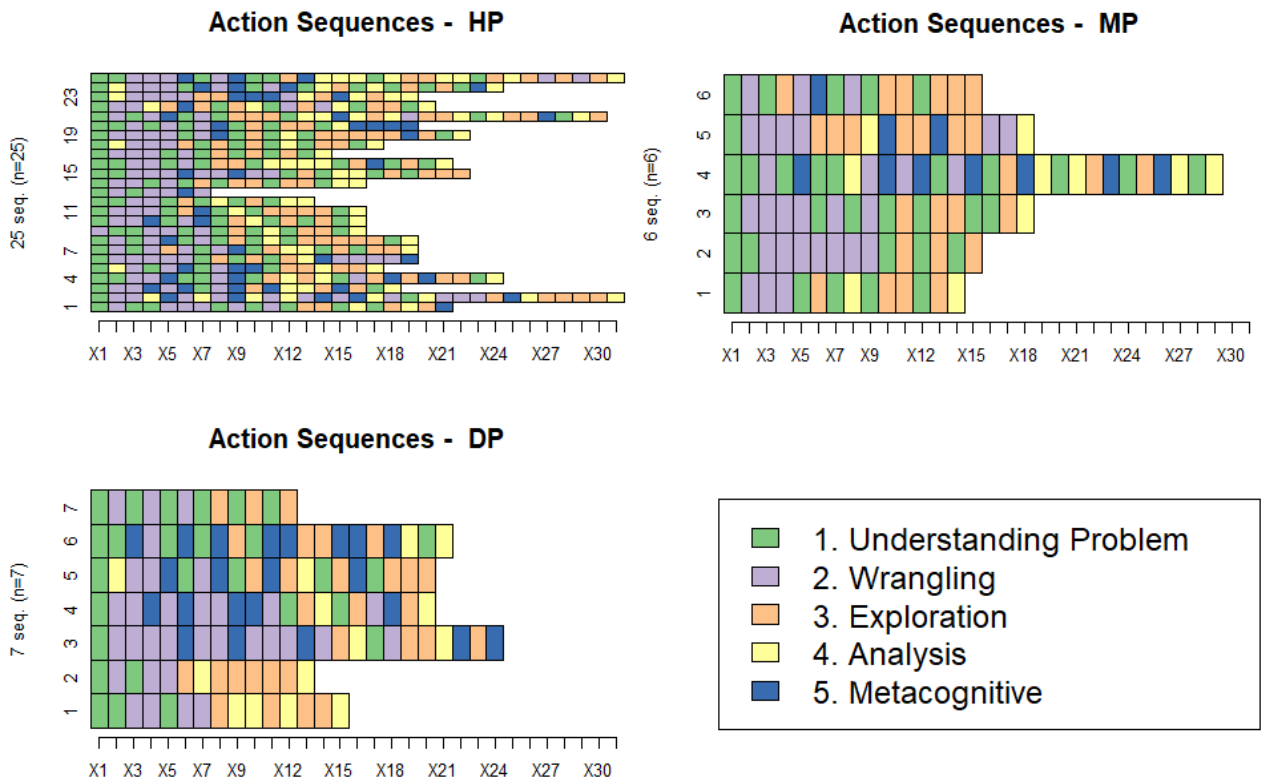

Figure 11. Sequence of actions per participant during the easy task, grouped by type of action to facilitate interpretation.

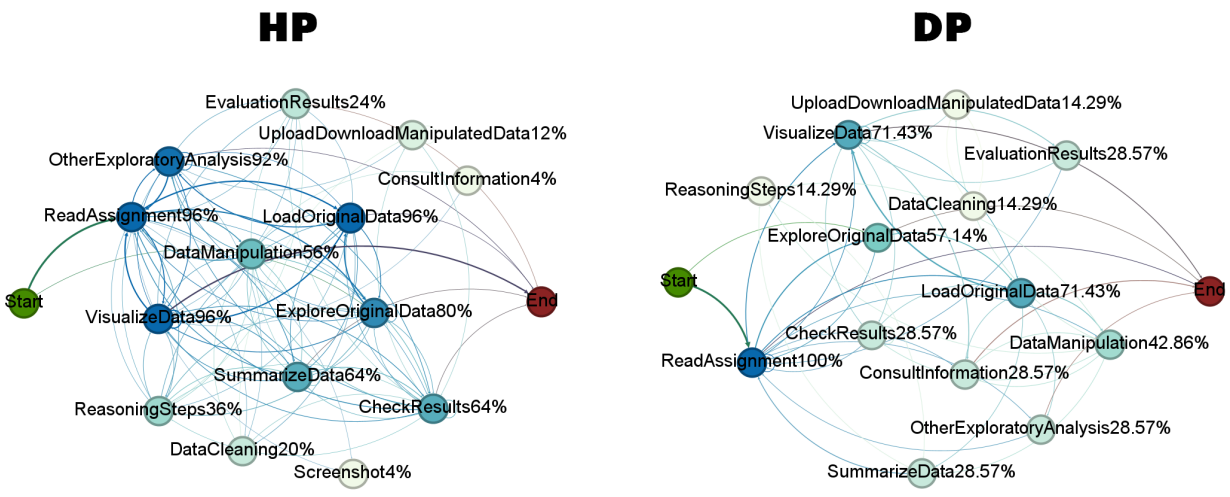

Figure 12. High-level overview of the process maps for the HP and DP performance clusters during the hard task. This figure highlights the number of distinct states and the connections between them. The nodes represent various actions performed by the participants, while the links indicate that one action was conducted immediately before another. The colour of each node reflects the percentage of participants performing that action (blue indicates a high percentage, while green represents a low percentage). The width of each link represents the percentage of times two actions occurred consecutively in the sequence.

First, we created an LMM that predicts the final score based on the sequence length (covariant) and complexity index

Table 6. Mean and standard deviation of sequence length and complexity for different performance clusters.

Cluster	Easy Task		Medium-Difficulty Task		Hard Task	
	Length *M* (*SD*)	Complexity *M* (*SD*)	Length *M* (*SD*)	Complexity *M* (*SD*)	Length *M* (*SD*)	Complexity *M* (*SD*)
HP	19.6 (5.5)	0.83 (0.08)	24.0 (7.2)	0.82 (0.07)	16.8 (6.6)	0.77 (0.85)
MP	18.2 (5.6)	0.75 (0.12)	21.7 (8.7)	0.78 (0.07)	15.3 (5.4)	0.83 (0.05)
DP	17.9 (4.5)	0.80 (0.08)	19.9 (7.6)	0.67 (0.13)	9.9 (7.2)	0.55 (0.4)

(covariant) factored by task difficulty and clustered by participant [Marginal $R^2 = 0.267$, LRT $\chi^2(6) = 36.575, p < 0.001$]. This model explains 26.7% of the variance in the task scores. In this model, only the complexity index is a statistically significant covariant [$F(1,104.7) = 11.801, p < 0.001$]. A simpler model including only complexity index and task difficulty [Adj. $R^2 = 0.218, F(3,108) = 7.29, p < 0.001$] is more parsimonious and not statistically less effective than the complete model [$\Delta R^2 = 0.028$, LRT $\chi^2(3) = 4.227, p = 0.238$]. In the smaller model, the complexity index is still a significant factor with a larger effect size [$F(1,102.9) = 15.416, p < 0.001$]. There is no statistically significant interaction between complexity index and task difficulty. These results indicate that sequence complexity alone accounts for a good portion of the variance in task scores.

Given that sequence complexity is an important factor for determining the task score, we will use it to help us determine if it can explain the membership of participants in different performance groups. A multinomial logistic regression using complexity index and task difficulty as IVs [$R^2_N = 0.0841, \chi^2(6,114) = 13.0, p = 0.043$] found a significant association between complexity index and the performance cluster [$\chi^2(2,114) = 12.998, p = 0.002$]. An exploration of the coefficients shows that participants with sequences of high complexity are much more likely to belong to the HP cluster than the DP cluster. No statistically significant relationships were found for complexity index between the HP and MP clusters.

4.5.2 Action Frequency

The descriptive statistics provided by the process map suggested that there were differences in the use of different actions between participants in each cluster to solve each task. To statistically test these differences, we calculated the frequency of use of each type of action.

An LMM factored by task difficulty, clustered by participant, and containing the frequency of each type of activity per participant per task as covariant IVs was created. The resulting model [marginal $R^2 = 0.342$, LRT $\chi^2(14) = 40.083, p < 0.001$] explains 34% of the variability of the score. In this model, checking results [$F(1,60.8) = 6.76, p = 0.012$], exploring original data [$F(1,60.9) = 6.49, p = 0.013$], and evaluating results [$F(1,59.4) = 6.72, p = 0.012$] were significant predictors of the test score. A more parsimonious model with just these variables performs similarly [$\Delta.R^2 = 0.154, LTR\chi^2(11) = 3.660, p < 0.976$], and the three actions continue to be significant predictors. An analysis of the coefficients connected to these variables shows checking results is positively related to scores [coef $= 0.272, p = 0.002$], exploring original data is inversely related to scores [coef $= -0.457, p = 0.022$], and evaluation of results is positively related to scores [coef $= 0.572, p < 0.001$]. Again, the information contained in the sequence of actions, in this case the frequency of each action in the sequence, is a good predictor of the participant's performance.

We also used the frequency of actions to try to explain the performance cluster membership. We applied a multinomial logistic regression using the frequency of each action as IV. The resulting model [$R^2_N = 0.496, \chi^2(30, N = 76) = 56.6, p = 0.002$] found that the statistically differentiating factor for participants in the DP and HP clusters is reading the assignment [odds ratio $= 0.510, p = 0.030$]. More importantly, this model found several statistically significant factors that differentiate MP from HP participants: checking results [odds ratio $= 0.561, p = 0.042$], evaluating results [odds ratio $= 0.312, p = 0.028$], and exploring original data [odds ratio $= 3.701, p = 0.017$]. It is important to note that exploring original data has an odds ratio greater than one, meaning that the higher the frequency of checking the original data, the higher the probability of being a member of the MP group. Finding similar variables in the explanation of the participants' membership in the performance cluster confirms their importance to better understand the scores obtained by the participants.

5. Discussion and Implications

Once we analyzed the data and obtained the results, we proceeded to discuss their meaning and implications for the LA research and practice communities. We follow the same order in which the analyses were conducted.

5.1 RQ1. Effect of Level of Expertise on Performance

In general, most participants (75%) scored 20 or higher out of a maximum of 30 points. This is surprising, as the different tasks, especially the medium and hard difficulty levels, require specialized data-cleaning, manipulation, analysis, and visualization

skills. While the experiment does not offer a control group, the pre-test serves as an indicator of baseline performance. Even participants who were not able to solve the most basic technical questions on their own in the pre-test were able to get the right answers to most of the questions in the experimental tasks. This was the first indication of the equalizing power that the use of ChatGPT could have for supporting non-experts in conducting descriptive LA practice.

Also interesting to note is that the average score in the hard task (network visualization) is very similar to the score in the medium-difficulty task. This was not expected, as the creation of a network visualization from a log file, the combination of different datasets, and data imputation in the hard question require complex data wrangling that is not immediately obvious, even for students of LA courses. On the other hand, the medium-difficulty task required the identification of errors and typos in the data that are technically very easy to correct if you can identify them. These results offer a more nuanced interpretation of the findings of previous work (Zheng, 2023; Shen et al., 2024; Maddigan & Susnjak, 2023; Owolabi et al., 2024). The results provided by ChatGPT do not necessarily get worse with the complexity of the task, as measured by the effort or technical knowledge required by a human to solve it. The correctness of ChatGPT also depends on the clarity and ambiguity with which the problem and the solution can be stated. The use of ChatGPT seems to reduce the technical complexity of the tasks, but it does not reduce the requirement of a certain level of domain knowledge and common sense.

The statistical analysis for RQ1 was conclusive. The short answer to this question is yes: the participants' level of previous technical expertise in solving descriptive data analysis tasks statistically influences their scores in descriptive LA tasks when using the support of ChatGPT v4 with Code Interpreter. The desired level of statistical power was achieved, and these results should be viewed as strong evidence that the role of previous expertise in data analysis does not become irrelevant with the use of ChatGPT. This aligns with previous research showing that expertise plays an important role in final performance (Zheng, 2023; Shen et al., 2024; Maddigan & Susnjak, 2023; Owolabi et al., 2024). However, our analysis reveals several caveats to this positive answer. First, the difference was only found between experts and novices. Intermediate participants are not statistically different from either of the other groups. Second, the difference in average scores between experts and novices is small (less than 3 points out of 30). And third, the IV level of expertise only explained 16% of the observed variability in the LMM to predict task scores.

Our full original hypothesis, that the difference should be significantly stronger for harder tasks, was not confirmed by the statistical analysis. While this is somewhat visible in the visualizations (see the right side of Figure 3), there was no statistically significant interaction between level of expertise and task difficulty to predict the task score. Contextualizing these results, while it was expected that the use of ChatGPT would help novices to perform better than expected, especially for the easy task, the similarity in final scores between expertise levels and the non-statistical difference in scores of the medium and hard tasks were positively surprising.

These positive results, however, do not imply that any teacher, right now, without knowledge of data science, could start using ChatGPT to obtain answers from data about their students. The scores obtained by novices were high but far from perfect. In real LA practice, at least on the technical side, performance is not measured by nuanced scores but by getting the data analysis right every time. The main insight that these results provide to the LA field is the great potential of LLMs to lower the barrier of access to the practice in the field. If the use of an unrefined version of ChatGPT could already drastically reduce the expertise gap, the design and creation of a fine-tuned or tailored version of an LLM that takes into account the educational context in which it is deployed, is prepared to address common pitfalls, reminds the user of necessary steps on its own, and could be integrated into learning management systems (LMSs) or student information systems (SISs) could revolutionize both the perception and adoption of LA.

5.2 RQ 2. Exploration of Other Explanatory Factors

The second part of our analysis focused on finding which other factors could also explain the variability of the scores. Given that the previous level of technical expertise in data analysis is not a good predictor of the score, we explored other variables related to other types of previous expertise; their previous confidence in being able to solve the tasks; the time invested per task; the number, length, and other characteristics of the prompts used; and, finally, the sequence of actions used by participants to solve the tasks.

The nature of these analyses is exploratory and different from the inferential analysis for RQ1. The number of participants was selected to address the first research question and the repeated-measures ANOVA analysis was used to answer it. While using similar statistical methods, the results of these analyses should be interpreted as indicators or initial leads to further investigate their relation to individuals' performance while using ChatGPT to solve LA tasks. Given this exploratory nature, there is important information not only in discovering that a factor is statistically significantly related to the score but also in finding those that are not.

The first example of the value of negative results comes from the analysis of variables related to previous experience with ChatGPT and confidence in being able to solve the tasks. The fact that differences in perceived proficiency in the use of ChatGPT, the frequency of its past use, and whether participants have used ChatGPT before to solve data analysis problems were not indicative factors of score implies that the use of the ChatGPT interface is so intuitive that one can become proficient

very rapidly, even without prior experience. The naturalness and ease of use of conversational interfaces, such as those presented by ChatGPT, are well known (Mulia et al., 2023). This offers a stark contrast with the difficulty and complexity of using and understanding dashboard-based LA applications (Park & Jo, 2019). The main implication of this for the LA field is that we should re-evaluate our preferred tool (Verbert et al., 2020) for communicating and interacting with LA information. The work being done in narrative-based LA (Fernandez Nieto et al., 2022) is a good step in this direction.

More interesting than the negative result for confidence level to predict final scores is how it is distributed among participants of different expertise. As can be seen in the rightmost graph of Figure 6, novices in data analysis were more confident than intermediate and expert users. This result reinforces the findings that individuals have a positive attitude toward the capabilities of LLMs to solve any type of cognitive task, especially if they are not experts in that task and are not aware of its inner complexities (Prather et al., 2023). This result is one of the first calls to caution in this work about the use of LLMs by non-experts to solve LA problems. Novices seem to compensate for their lack of expertise with an over-reliance on the capabilities of GenAI systems, in a clear example of automation bias (Goddard et al., 2012). If the LA community wishes to build LLM-based systems to help non-experts conduct their own analyses, these systems should be programmed to offer automatic sanity checks, such as the use of self-correcting LLMs (Pan et al., 2024).

The exploration of time- and prompt-related characteristics also provided important insights. The most important is that the originality and sophistication of the language used were not significant predictors of performance. Participants who expressed themselves using data science technical terms instead of colloquial equivalents did not perform better than their peers. While surprising in this study, as we expected some level of significance, this finding is justified given the demonstrated capabilities of LLMs to understand and process natural language, even if it contains errors or inaccuracies (Cecchini et al., 2024). The other important finding in this group of variables was that total length is a statistically significant predictor of higher scores. One possible way to interpret these results is that the more information provided to the LLM, the better job it will do in arriving at the correct answer. This is a well-known phenomenon with LLMs that is exploited by several prompt-engineering techniques (Marvin et al., 2023). The main implication of the results in this section for the LA community, unsurprisingly, is that how we divide and specify our prompts is more important than the actual language used. As mentioned before, prompt-engineering techniques are still a major part of the successful use of LLMs.

The time on task and prompt-related variables are also the first significant predictors of the performance cluster assigned to participants based on their scores in different tasks. The number of prompts differentiates between high and middle performers, while total length differentiates between high and declining performers. The difference in the effect of these two related but not redundant variables on the performance of participants is not completely clear from the conducted analyses. Based on the behaviour of the performance clusters (see Figure 4), the lack of co-significance between these variables, and the results found in other analyses, we hypothesize that using more prompts is related to overall higher performance, while the total length of the prompts affects only the more difficult tasks. Dividing a large prompt into several smaller but more concise prompts is one of the strategies suggested to get better results with LLMs (Marvin et al., 2023). On the other hand, complex problems that could be interpreted in different ways benefit from additional context and explanations to the LLM. More research needs to be conducted to establish if this hypothesis is indeed true.

Our original hypothesis for RQ2 was that most of the explanatory variables would be contained in the sequence of actions performed by each participant to solve the task. This hypothesis was confirmed by the exploratory results. The complexity and frequency of the diverse actions all created high-performing models that explained around 26–34% of the variance in the participants' scores. The most surprising result in this group was the performance of a single measurement: the sequence complexity index. Participants who conducted more varied, but not necessarily larger, sequences obtained higher scores. Not as surprising was the finding that the actual actions performed during the sequence are good indicators of performance. The specific actions that were found significant shed light on an important insight: the actions that happened after using ChatGPT to solve a problem (checking results and evaluating results) were more significant positive predictors of performance than the actions conducted inside ChatGPT (data manipulation, summarization, and visualization, among others). These results align with the findings of studies in the literature review (Zheng, 2023; Shen et al., 2024; Maddigan & Susnjak, 2023; Owolabi et al., 2024) that unanimously highlighted the importance of interpreting and checking the answers provided by the LLM. Due to ChatGPT generating linguistically plausible but not necessarily correct answers, checking and evaluating these answers is the key differentiator between high and medium performers. This could be the most important but also most anticipated result of the analysis. Its main implication for the LA community is that LLMs have great potential for extending access to LA practice, but users, even non-experts, should have the necessary knowledge not to conduct the procedure themselves but at least to be able to check if the result is correct and interpret this result in the pedagogical context. In other words, LLMs could replace technical knowledge but not domain knowledge or common sense.

5.3 General Reflection

As a final reflection in this discussion of the results, we want to highlight the significance of what this work has found beyond models, charts, and p-values: the use of a general-purpose GenAI enabled non-experts in data science to conduct even hard

descriptive LA tasks at a similar level of performance as experts. Even if the effect is not universal, there are still measurable differences and many possible pitfalls. These results open the door to a new paradigm of LA practice and a whole new research avenue in the field. This new way of doing LA is bound to redefine what we teach our students in the classroom, what we build for our stakeholders, and how the educational community perceives LA. We hope that the LA research and practice communities seize this opportunity.

6. Limitations and Further Work

All experiments have limitations, and ours has its fair share. While some of them are in the analysis we conducted, there are several that are precisely those analyses that we were not able to include. This section will describe these limitations, how they affect the results, and further research to address them.

- **Low ecological validity:** The tasks were arbitrarily assigned to the participants as they were not directly interested in solving them. The extrinsic motivation of compensation cannot be compared to the internal motivation of answering a self-generated question about their own practice. While this obviously affects the results, it does not invalidate the main findings. Self-motivated novice teachers are expected to apply more, not less, effort and time in solving the tasks. Further studies should be conducted with real practitioners.
- **Generalizability of results:** While our study's sample size was determined to be sufficient based on statistical power calculations, a key limitation lies in the lack of diversity within both the participant pool (e.g., they belong to the same geographical location and have similar level of access to technology) and the tasks assessed (e.g., the tasks did not cover all possible descriptive analytics skills). The homogeneity of our sample may limit the generalizability of our findings to broader populations. Additionally, the tasks selected for this study may not fully capture the variability present in real-world scenarios. Future research should aim to include more diverse participants and a wider range of tasks to enhance the robustness of the results.
- **Lack of control condition:** Having a control condition was not feasible for this study. Given the study's design, this work does not make any general claims about the impact of the experimental condition in comparison to its absence. The main research questions of the paper instead focus on differences between individuals with varying skill levels when LLM support is available. To test the effect of the LLM support on each individual, further studies could have a control condition where participants interact with a human LA expert to solve the same tasks.
- **Methodological limitations:** The experiment was only sized and designed to fully answer the first research question. The number of variables that we were able to safely include in the exploratory models was limited by the sample size. Also, the complexity of the action sequence data reduced the availability of suitable methods of analysis that go beyond description. Being aware of this limitation, we clearly stated in our results and discussion which analyses were inferential and which ones were exploratory in nature.
- **Lack of qualitative analysis:** While qualitative data was collected in the experiment (final interview), we decided not to include its analysis in the present work. The inclusion of the qualitative analysis of participants' experiences and its interaction with the quantitative results would have doubled the length of what is already a long manuscript. Several insights revealed by this data deserve to be discussed in depth in future work.
- **Learning evaluation skills:** One of the main unresolved questions raised by this work is that even non-experts need to have some level of knowledge—not necessarily to solve the data analysis tasks independently but at least to understand the LLM's output and evaluate whether it is accurate and aligned with their needs. Is this set of skills distinct from those required to solve the problem initially? How can these skills be taught or learned? These are intriguing questions for future research.
- **Ethical implications:** Asserting that LLMs can be used to solve descriptive LA tasks is very different from asserting that it is beneficial to do so. The results of this study should be viewed within the broader discussion of LLMs' impact on human agency (Darvishi et al., 2024) and creativity (Habib et al., 2024), as well as practical considerations like student data privacy (Yao et al., 2024) and accountability for errors (Santoni de Sio & Mecacci, 2021). We acknowledge that by focusing on very specific research questions, we have not addressed these larger issues. We encourage researchers, including our future selves, to explore these topics using both theoretical and empirical approaches.

The analyses conducted in this work only grazed the possible information contained in the sequence of actions of participants to better understand how ChatGPT or equivalent LLMs could be used to facilitate LA practice. Further studies, expressly designed to study this type of data in detail, could help us to better understand how the technical side of LA practice is conducted and how to better support it, as well as the ethical and practical implications of doing so.

7. Conclusions

This study explored the feasibility of using GenAI, specifically ChatGPT, to facilitate access to LA practice by enabling non-experts to conduct descriptive data analysis tasks. The findings reveal that while previous technical expertise in data analysis does influence performance, the difference between experts and novices is not as substantial as anticipated. This indicates that ChatGPT can significantly lower the barrier to entry for LA, making it accessible to a broader range of stakeholders, including those without a background in data science. Despite the promising results, the study also highlights that the successful use of ChatGPT for LA requires domain knowledge, the ability to formulate clear and unambiguous questions, and the ability to evaluate and check the answers provided by the model.

Given the rapid advancement of LLMs and the evolving familiarity and literacy of stakeholders with AI systems, the results of this work are likely to become obsolete in a short time. However, the primary purpose of these findings is to serve as a conversation starter within the LA community on how to leverage the capabilities of GenAI to scale and expand the practice of LA.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from New York University in the form of a research grant.

References

Ahn, S. (2024). Data science through natural language with ChatGPT's Code Interpreter. *Translational and Clinical Pharmacology*, *32*(2), 73–82. https://doi.org/10.12793/tcp.2024.32.e8

Campos, F. C., Ahn, J., DiGiacomo, D. K., Nguyen, H., & Hays, M. (2021). Making sense of sensemaking: Understanding how K–12 teachers and coaches react to visual analytics. *Journal of Learning Analytics*, *8*(3), 60–80. https://doi.org/10.18608/jla.2021.7113

Carlisle, S. (2018). Software: Tableau and Microsoft Power BI. *Technology||Architecture + Design*, *2*(2), 256–259. https://doi.org/10.1080/24751448.2018.1497381

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. *Physica Medica*, *83*, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

Cecchini, D., Nazir, A., Chakravarthy, K., & Kocaman, V. (2024). Holistic evaluation of large language models: Assessing robustness, accuracy, and toxicity for real-world applications. In A. Ovalle, K.-W. Chang, Y. T. Cao, N. Mehrabi, J. Zhao, A. Galstyan, J. Dhamala, A. Kumar, & R. Gupta (Eds.), *Proceedings of the Fourth Workshop on Trustworthy Natural Language Processing* (TrustNLP 2024), 21 June 2024, Mexico City, Mexico (pp. 109–117). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.trustnlp-1.11

Ceruzzi, P. (1996). From scientific instrument to everyday appliance: The emergence of personal computers, 1970–77. *History and Technology, an International Journal*, *13*(1), 1–31. https://doi.org/10.1080/07341519608581893

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In *Proceedings of the Second International Conference on Learning Analytics and Knowledge* (LAK 2012), 29 April–2 May 2012, Vancouver, British Columbia, Canada (pp. 134–138). ACM. https://doi.org/10.1145/2330601.2330636

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213. https://doi.org/10.1037/h0026256

Daele, S. V., & Janssenswillen, G. (2022). Identifying the steps in an exploratory data analysis: A process-oriented approach. In M. Montali, A. Senderovich, & M. Weidlich (Eds.), *Process mining workshops. ICPM 2022. Lecture notes in business information processing* (pp. 526–538, Vol. 468). Springer. https://doi.org/10.1007/978-3-031-27815-0_38

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*, *210*, 104967. https://doi.org/10.1016/j.compedu.2023.104967

Dourado, R. A., Rodrigues, R. L., Ferreira, N., Mello, R. F., Gomes, A. S., & Verbert, K. (2021). A teacher-facing learning analytics dashboard for process-oriented feedback in online learning. In *Proceedings of the 11th International Conference on Learning Analytics and Knowledge* (LAK 2021), 12–16 April 2021, Irvine, California, USA (pp. 482–489). ACM. https://doi.org/10.1145/3448139.3448187

Echeverria, V., Martinez-Maldonado, R., Buckingham Shum, S., Chiluiza, K., Granda, R., & Conati, C. (2018). Exploratory versus explanatory visual learning analytics: Driving teachers' attention through educational data storytelling. *Journal of Learning Analytics*, *5*(3), 73–97. https://doi.org/10.18608/jla.2018.53.6

Echeverria, V., Yan, L., Zhao, L., Abel, S., Alfredo, R., Dix, S., Jaggard, H., Wotherspoon, R., Osborne, A., Buckingham Shum, S., Gasevic, D., & Martinez-Maldonado, R. (2024). Teamslides: A multimodal teamwork analytics dashboard for teacher-guided reflection in a physical learning space. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 112–122). ACM. https://doi.org/10.1145/3636555.3636857

Estrellado, R., Freer, E., Rosenberg, J., & Velásquez, I. (2020). *Data science in education using R*. Routledge. https://doi.org/10.4324/9780367822842

Fernandez Nieto, G. M., Kitto, K., Buckingham Shum, S., & Martinez-Maldonado, R. (2022). Beyond the learning analytics dashboard: Alternative ways to communicate student data insights combining visualisation, narrative and storytelling. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 219–229). ACM. https://doi.org/10.1145/3506860.3506895

Field, A. P. (2005). Kendall's coefficient of concordance. In B. Everitt & D. Powell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1010–1011, Vol. 2). Wiley. https://doi.org/10.1002/0470013192.bsa327

Flanagan, B., Wasson, B., & Gašević, D. (Eds.). (2024). *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan. ACM. https://doi.org/10.1145/3636555

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378. https://doi.org/10.1037/h0031619

Furche, T., Gottlob, G., Libkin, L., Orsi, G., & Paton, N. W. (2016). Data wrangling for big data: Challenges and opportunities. In W. Martens & T. Zeume (Eds.), *Proceedings of the 19th International Conference on Extending Database Technology* (ICDT 2016), 15–18 March 2016, Bordeaux, France (pp. 473–478). Schloss Dagstuhl—Leibniz-Zentrum für Informatik. https://doi.org/10.5441/002/edbt.2016.44

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goswami, A., Ramakrishna, G., & Sethi, R. (2021). Review on smile detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *7*(2), 577–583. https://doi.org/10.32628/CSEIT2172134

Gundlach, E., & Ward, M. D. (2021). The data mine: Enabling data science across the curriculum. *Journal of Statistics and Data Science Education*, *29*(sup1), S74–S82. https://doi.org/10.1080/10691898.2020.1848484

Habib, S., Vogel, T., Anli, X., & Thorne, E. (2024). How does generative artificial intelligence impact student creativity? *Journal of Creativity*, *34*(1), 100072. https://doi.org/10.1016/j.yjoc.2023.100072

Holstein, K., Hong, G., Tegene, M., McLaren, B. M., & Aleven, V. (2018). The classroom as a dashboard: Co-designing wearable cognitive augmentation for K-12 teachers. In *Proceedings of the Eighth International Conference on Learning Analytics and Knowledge* (LAK 2018), 7–9 March 2018, Sydney, Australia (pp. 79–88). ACM. https://doi.org/10.1145/3170358.3170377

Jaimovitch-López, G., Ferri, C., Hernández-Orallo, J., Martínez-Plumed, F., & Ramírez-Quintana, M. J. (2023). Can language models automate data wrangling? *Machine Learning*, *112*(6), 2053–2082. https://doi.org/10.1007/s10994-022-06259-9

Kaliisa, R., & Dolonen, J. A. (2023). Cada: A teacher-facing learning analytics dashboard to foster teachers' awareness of students' participation and discourse patterns in online discussions. *Technology, Knowledge and Learning*, *28*(3), 937–958. https://doi.org/10.1007/s10758-022-09598-7

Khosravi, H., Viberg, O., Kovanovic, V., & Ferguson, R. (2023). Generative AI and learning analytics. *Journal of Learning Analytics*, *10*(3), 1–6. https://doi.org/10.18608/jla.2023.8333

Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*, *25*(6), 1066–1078. https://doi.org/10.1002/cae.21844

Leiner, B. M., Cerf, V. G., Clark, D. D., Kahn, R. E., Kleinrock, L., Lynch, D. C., Postel, J., Roberts, L. G., & Wolff, S. (2009). A brief history of the internet. *ACM SIGCOMM Computer Communication Review*, *39*(5), 22–31. https://doi.org/10.1145/1629607.1629613

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410. https://doi.org/10.3390/educsci13040410

Maddigan, P., & Susnjak, T. (2023). Chat2VIS: Generating data visualizations via natural language using ChatGPT, Codex and GPT-3 large language models. *IEEE Access*, *11*, 45181–45193. https://doi.org/10.1109/ACCESS.2023.3274199

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, *60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., & Buckingham Shum, S. (2020). From data to insights: A layered storytelling approach for multimodal learning analytics. In *Proceedings of the 2020 CHI Conference on*

Human Factors in Computing Systems (CHI 2020), 25–30 April 2020, Honolulu, Hawaii, USA (pp. 1–15). ACM. https://doi.org/10.1145/3313831.3376148

Martinez-Maldonado, R., Elliott, D., Axisa, C., Power, T., Echeverria, V., & Buckingham Shum, S. (2022). Designing translucent learning analytics with teachers: An elicitation process. *Interactive Learning Environments*, *30*(6), 1077–1091. https://doi.org/10.1080/10494820.2019.1710541

Martinez-Maldonado, R., Kay, J., Yacef, K., & Schwendimann, B. (2012). An interactive teacher's dashboard for monitoring groups in a multi-tabletop learning environment. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems. ITS 2012. Lecture notes in computer science* (pp. 482–492, Vol. 7315). Springer. https://doi.org/10.1007/978-3-642-30950-2_62

Marvin, G., Hellen, N., Jjingo, D., & Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In I. Jacob, S. Piramuthu, & P. Falkowski-Gilski (Eds.), *Data intelligence and cognitive informatics. ICDICI 2023. Algorithms for intelligent systems* (pp. 387–402). Springer. https://doi.org/10.1007/978-981-99-7962-2_30

McGrath, A. L. (2014). Content, affective, and behavioral challenges to learning: Students' experiences learning statistics. *International Journal for the Scholarship of Teaching and Learning*, *8*(2), 6. https://doi.org/10.20429/ijsotl.2014.080206

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org//10.11613/BM.2012.031

Mulia, A. P., Piri, P. R., & Tho, C. (2023). Usability analysis of text generation by ChatGPT OpenAI using system usability scale method. *Procedia Computer Science*, *227*, 381–388. https://doi.org/10.1016/j.procs.2023.10.537

Ochoa, X. (2022). Multimodal learning analytics: Rationale, process, examples, and direction. In C. Lang, G. Siemens, A. Wise, D. Gasevic, & A. Merceron (Eds.), *The handbook of learning analytics* (pp. 54–65). SoLAR. https://doi.org/10.18608/hla22.006

Owolabi, A. T., Okunlola, O. O., Adewuyi, E. T., Idowu, J. I., & Oladapo, O. J. (2024). The advent of ChatGPT: Job made easy or job loss to data analysts. *WSEAS Transactions on Computers*, *23*, 24–40. https://doi.org/10.37394/23205.2024.23.3

Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, *12*, 484–506. https://doi.org/10.1162/tacl_a_00660

Park, Y., & Jo, I.-H. (2019). Factors that affect the success of learning analytics dashboards. *Educational Technology Research and Development*, *67*(6), 1547–1571. https://doi.org/10.1007/s11423-019-09693-0

Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics—A systematic review. *Education and Information Technologies*, 1–30. https://doi.org/10.1007/s10639-023-12401-4

Plakandaras, V., Gogas, P., Papadimitriou, T., & Tsamardinos, I. (2022). Credit card fraud detection with automated machine learning systems. *Applied Artificial Intelligence*, *36*(1), 2086354. https://doi.org/10.1080/08839514.2022.2086354

Pozdniakov, S., Martinez-Maldonado, R., Tsai, Y.-S., Echeverria, V., Srivastava, N., & Gasevic, D. (2023). How do teachers use dashboards enhanced with data storytelling elements according to their data visualisation literacy skills? In *Proceedings of the 13th International Conference on Learning Analytics and Knowledge* (LAK 2023), 13–17 March 2023, Arlington, Texas, USA (pp. 89–99). ACM. https://doi.org/10.1145/3576050.3576063

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2023). "It's weird that it knows what I want": Usability and interactions with Copilot for novice programmers. *ACM Transactions on Computer-Human Interaction*, *31*(1), 1–31. https://doi.org/10.1145/3617367

Raja, K., & Ramathilagam, S. (2021). Washing machine using fuzzy logic controller to provide wash quality. *Soft Computing*, *25*(15), 9957–9965. https://doi.org/10.1007/s00500-020-05477-4

Ritschard, G. (2023). Measuring the nature of individual sequences. *Sociological Methods & Research*, *52*(4), 2016–2049. https://doi.org/10.1177/00491241211036156

Sá, D., Guimarães, T., Abelha, A., & Santos, M. F. (2024). Low code approach for business analytics. *Procedia Computer Science*, *231*, 421–426. https://doi.org/10.1016/j.procs.2023.12.228

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Sarmiento, J. P., & Wise, A. F. (2022). Participatory and co-design of learning analytics: An initial review of the literature. In *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK 2022), 21–25 March 2022, online (pp. 535–541). ACM. https://doi.org/10.1145/3506860.3506910

Sharma, A. K., Sharma, D. M., Purohit, N., Rout, S. K., & Sharma, S. A. (2022). Analytics techniques: Descriptive analytics, predictive analytics, and prescriptive analytics. In P. Jeyanthi, T. Choudhury, D. Hack-Polay, T. Singh, & S. Abujar

(Eds.), *Decision intelligence analytics and the implementation of strategic business management. EAI/Springer innovations in communication and computing* (pp. 1–14). Springer. https://doi.org/10.1007/978-3-030-82763-2_1

Shen, Y., Ai, X., Soosai Raj, A. G., Leo John, R. J., & Syamkumar, M. (2024). Implications of ChatGPT for data science education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (SIGCSE 2024), 20–23 March 2024, Portland, Oregon, USA (pp. 1230–1236). ACM. https://doi.org/10.1145/3626252.3630874

Shreiner, T. L., & Dykes, B. (2021). Visualizing the teaching of data visualizations in social studies: A study of teachers' data literacy practices, beliefs, and knowledge. *Theory & Research in Social Education*, *49*(2), 262–306. https://doi.org/10.1080/00933104.2020.1850382

Singh, A., Singh, S., & Rathee, J. (2023). Basic principles of data wrangling. In M. Niranjanamurthy, K. Sheoran, G. Dhand, & P. Kaur (Eds.), *Data wrangling: Concepts, applications and tools* (pp. 1–18). Wiley. https://doi.org/10.1002/9781119879862.ch1

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(1), 214–216. https://doi.org/10.1038/d41586-023-00340-6

Sullivan, A. (2014). *Determining an inter-rater agreement metric for researchers evaluating student pathways in problem solving.* [Master's thesis, Iowa State University, Iowa]. https://doi.org/10.31274/etd-180810-2820

Toosi, A., Bottino, A. G., Saboury, B., Siegel, E., & Rahmim, A. (2021). A brief history of AI: how to prevent another winter (a critical review). *PET Clinics*, *16*(4), 449–469. https://doi.org/10.1016/j.cpet.2021.07.001

van Leeuwen, A., Knoop-van Campen, C. A., Molenaar, I., & Rummel, N. (2021). How teacher characteristics relate to how teachers use dashboards: Results from two case studies in K-12. *Journal of Learning Analytics*, *8*(2), 6–21. https://doi.org/10.18608/jla.2021.7325

van Leeuwen, A., Teasley, S. D., & Wise, A. F. (2022). Teacher and student facing learning analytics. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *Handbook of learning analytics* (2nd edition, pp. 130–140). SoLAR. https://doi.org/10.18608/hla22.013

Verbert, K., Ochoa, X., De Croon, R., Dourado, R. A., & De Laet, T. (2020). Learning analytics dashboards: The past, the present and the future. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge* (LAK 2020), 23–27 March 2020, Frankfurt, Germany (pp. 35–40). ACM. https://doi.org/10.1145/3375462.3375504

Wallace, D., & Green, S. B. (2013). Analysis of repeated measures designs with linear mixed models. In D. S. Moskowitz & S. L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data* (pp. 103–134). Psychology Press. https://psycnet.apa.org/record/2001-05300-005

Wise, A. F. (2014). Designing pedagogical interventions to support student use of learning analytics. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (LAK 2014), 24–28 March 2014, Indianapolis, Indiana, USA (pp. 203–211). ACM. https://doi.org/10.1145/2567574.2567588

Wise, A. F., & Jung, Y. (2019). Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, *6*(2), 53–69. https://doi.org/10.18608/jla.2019.62.4

Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *Proceedings of the 14th International Conference on Learning Analytics and Knowledge* (LAK 2024), 18–22 March 2024, Tokyo, Japan (pp. 101–111). ACM. https://doi.org/10.1145/3636555.3636856

Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, *4*(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Zheng, Y. (2023). ChatGPT for teaching and learning: An experience from data science education. In *Proceedings of the 24th Annual Conference on Information Technology Education* (SIGITE 2023), 11–14 October, Marietta, Georgia, USA (pp. 66–72). ACM. https://doi.org/10.1145/3585059.3611431

Zhu, J.-P., Cai, P., Niu, B., Ni, Z., Xu, K., Huang, J., Wan, J., Ma, S., Wang, B., Zhang, D., Tang, L., & Liu, Q. (2024). Chat2Query: A zero-shot automatic exploratory data analysis system with large language models. In *2024 IEEE 40th International Conference on Data Engineering* (ICDE 2024), 13–16 May 2024, Utrecht, Netherlands (pp. 5429–5432). IEEE. https://doi.org/10.1109/ICDE60146.2024.00420