Augmenting Active Learning with GenAI: Enhancement or Impairment? Evidence from a Data Visualization Course

Vladimir Macko[1], Abdessalam Ouaazki[1], Natalia Bartłomiejczyk[1], Barbara Monteiro[1], Pascal Felber[1], and Adrian Holzer[1,2]

[1] University of Neuchâtel, Rue A.-L. Breguet 2, 2000 Neuchâtel, Switzerland
{vladimir.macko, abdessalam.ouaazki, natalia.bartlomiejczyk, barbara.monteiro, pascal.felber,adrian.holzer}@unine.ch
[2] TD School, University of Technology Sydney, Ultimo NSW 2007, Australia

Abstract. As Generative Artificial Intelligence (GenAI) is becoming ubiquitous across learning domains, it is crucial to better understand how learning experiences could take advantage of its possibilities and avoid its pitfalls. In this paper, we address this issue by focusing on the context of a data visualization course. What makes this context unique is its combination of two areas where GenAI has shown notable effectiveness: writing code and storytelling. To evaluate how undergraduate students would leverage GenAI in this context, we conducted an in-class between-subjects experiment (N=43) with a control (no GenAI) and treatment group (with GenAI). In the 60-minute experiment, students from the data visualization course were asked to prepare a data story within Jupyter Notebook, including both textual story elements and data visualization. In addition to these two groups, we included AI-only group in which task instructions were given directly to a GenAI tool without further human intervention. The results of our experiment indicate that students perceive GenAI as a tool improving both their learning experience and outcomes. However, an analysis of the learning outcomes exhibits no statistically significant difference between creations of students with or without GenAI support. Interestingly, the outputs generated by GenAI alone outperformed those of both student groups.

Keywords: Generative AI · Data Visualization · Jupyter Notebooks

1 Introduction

Generative Artificial Intelligence (GenAI) has taken over the world in the short two years since the public release of ChatGPT. The fields of learning and education are no exception, with GenAI affecting everything from assessment and authorship to learning tools and tutoring.

Even before GenAI, chatbots have already been investigated for a while in the learning context, as they can help with everything from content integration and access to increasing motivation, maximizing student abilities and engagement [2], providing rapid assistance [23], and saving time overall [31]. These chatbots

have become increasingly popular in education, largely because of their ability to improve the learning experience [32]. They simulate human conversation, allowing students to interact through natural language. These tools can make it easier to access learning materials, boost motivation and engagement, support large numbers of users at once, and provide instant help when needed [23]. They are also used as virtual assistants, helping with academic or administrative tasks [37], or answering common questions to create a more personalized and helpful experience [28,35,31].

With the rise of GenAI, a multitude of studies have investigated how this technology can be used in various computational thinking contexts such as Data-Oriented Programming, Python for Data Analysis, and Computational Thinking [29,46,8,26,25,17]. Other studies investigate AI's role in a variety of broad IT and programming-related courses, such as database management, text mining, computer networks and software testing [9,34,11,20,44,18,36]. GenAI allows for the creation of customized learning experiences tailored to individual students' needs, leading to improved learning outcomes [38,24]. It elevates efficiency and productivity by enabling greater task automation and fostering co-creation between humans and AI [39,27]. By handling routine tasks and content delivery, it also frees up educators to concentrate on higher-order thinking skills and provide more meaningful mentorship [6].

Data science is particularly impacted by GenAI as most coders[3] in the industry use GenAI tools as part of their workflows [14]. Some authors see the future as moving from writing code to prompting AI to write code for them [42]. Coders will focus more on quality control rather than syntactic errors [4]. As GenAI seems particularly well suited both for writing code and writing stories [4,45], it could potentially be leveraged to make it easier to craft such data stories [43] and make this task more accessible for less technically oriented individuals [1]. As such, it is crucial to better understand how learners can take advantage of the possibilities of GenAI to craft data stories and avoid their pitfalls [7]. In this paper, we tackle this issue by addressing the following question:

RQ: How can GenAI tools support learners in crafting data visualization stories?

To answer this question, this paper is structured as follows: Section 2 presents the related work, Section 3 the methodology, and Section 4 the results. Then Section 5 discusses the findings before Section 6 wraps up with a conlusion.

2 Related work

Several studies specifically investigated the use of GenAI tools for data visualization in educational contexts [41,21,15,5,3].

In terms of methodology, these studies range from a prospective critical analysis of a single author ($N = 1$) [3], to interview studies of several participants ($N = 10$) [5], to multimodal studies with repeated measure experiments in class

[3] `https://survey.stackoverflow.co/2023/#section-sentiment-and-usage-ai-tools-in-the-development-process`

($N = 59$ and $N = 26$) [41,15] or in the lab ($N = 39$) [21]. In these studies, the focus was generally solely on coding visualizations, rather than on the combination of storytelling with coding. Most studies did not investigate different conditions comparing a GenAI-augmented learning experience with a non-augmented one. Nonetheless, one study conducted a within-subjects experiment where students experienced several conditions with and without GenAI in the classroom [41].

In terms of findings, the studies explored how students used GenAI when they had access to it. They established that participants predominantly used it for coding assistance [15]. Reported benefits included both efficiency-related improvements, such as time savings, and enhanced learning outcomes [15,41]. Moreover, previous results suggested that GenAI could play an important role in broadening data analysis access as it was found to increase confidence [15] and lower the technical threshold for creating visualizations and conducting exploratory data analysis [3,21]. Most studies found positive user feedback [41,21]. Efficiency and lower access barriers could encourage creativity, as they allow for rapid prototyping and iterative refinement [5]. Among the identified risks, research mentions inaccuracies in GenAI outputs, difficulty in effectively prompting the tool, as well as concerns about over-reliance [15,3].

These studies provide promising preliminary findings and hint at open research gaps. First, while some studies included data narratives or storytelling in their task description, it is not clear how GenAI improves these aspects. Second, it is still unclear how a GenAI-augmented learning experience compares quantitatively with a learning experience without GenAI in terms of both students' perceived task efficiency and actual performance. This paper will aim to address these gaps by (1) setting up a task combining storytelling and data analysis and (2) conducting a between-subject experiment in the wild to better understand outcomes quantitatively.

3 Methodology

The study takes an experimental design approach in the wild. As such, we designed a controlled experiment with a treatment group in which students had access to GenAI (HAI, where H stands for humans) and a control group in which students were alone (H). This design allows for a between-group comparison. We gathered data in a multi-modal fashion, combining student surveys with learning outcomes and log data from chatbot conversations. In addition to these two groups, we also created a third group (AI) in which GenAI operated alone, based only on an initial human prompt, to investigate how it approaches the task with minimal human intervention.

3.1 Pedagogical scenario

As part of their final assignment for an undergraduate Data Visualization course, students were asked to prepare a Jupyter Notebook file containing annotated data visualizations in the context of a charity donation request. The dataset, sourced from `https://www.gapminder.org`, includes yearly values (2000–2022) for four indicators–access to water, sanitation, internet, and extreme poverty–across a selection of developing countries. The selected developing countries

were intentionally chosen in order to showcase the continuous improvement of the measured variables. To highlight a stark contrast, we also included a non-developing country with the highest Human Development Index in 2024, offering a benchmark against the data from developing countries. Figure 1 shows the 60-minute assignment description.

```
- Assignment -- You are a data visualization expert, and you have received a
  request from a local charity to prepare convincing support material for
  their campaign. Leaflets will be designed later on in the campaign based
  on your essential input. You should now:

- Prepare a simple but complete Jupyter Notebook file (.ipynb) that contains
  a captivating title, two commented, professionally looking Python
  visualizations using the provided data and a catchy punchline at the end.
  Limit the total text amount to 2 sentences per visualization at maximum.
  The notebook should contain all the Python code needed for their
  construction (the notebook is self-sufficient).

- Also, include a textual description of a suggestion of 1 illustrative
  photograph that should be included to enhance the impact of your
  visualizations (do not include the image, just describe how it should
  look). No other files should be prepared.

- Keep in mind that the main goal is to use storytelling to convince the
  general public that their contributions to charity make a real difference,
  and therefore, they should contribute. Only the data from the provided
  dataset (CharityDataset.csv) can be used in the visualizations, and no
  other sources should be included.

- Markdown cells can be used in the Jupyter notebook to include text
  comments. The text comments should be factually correct and should be
  related to the visualization presented. All visualizations should respect
  all the Visual design principles, and all good practices of data
  visualization should be followed.}
```

Fig. 1. Study assigmnent description

3.2 Procedure

The experiment was conducted during the final class of the Data Visualization course taught in French. The participants were bachelor students in data science, economics, and digital humanities. Out of 54 students taking this class, 46 students voluntarily participated in this study. They received a financial compensation equivalent to USD 10.

At the beginning of the class, students were randomly assigned to two groups based on the numbers they drew from an envelope. Groups were then placed in two different classrooms. All participants were presented with the same assignment description. This assignment was scored for study purposes but did not impact the students' grades for the course. The only difference between the groups was that group HAI was encouraged to use a GenAI tool such as ChatGPT or Copilot to help them to solve the task, while group H was explicitly asked not to use any AI tool and use only their knowledge and internet search engines to solve the tasks. Furthermore, group AI was generated by submitting the assignment description (in Figure 1) to ChatGPT4-o. To ensure coherence with the student notebooks, the following sentence was added at the end of the description: *Make sure you produce a Python Jupyter notebook with content in French that can be run.* We prompted the model 50 times, resulting in 50 different Jupyter Notebook files. We removed files that did not meet the expected

format (.ipynb) or those that could not be opened (24 occurrences in total). Accepted files ($N = 26$) from the AI group were then renamed for anonymity before they were scored. The study was approved by our ethics board.

3.3 Measures

We gathered data from three main sources: surveys to inquire about demographics and to measure student attitudes about the learning process, task output inspection to measure learning outcomes, and chatbot conversations to measure learning process behaviour for students in the HAI group. We also measured the level of academic performance by collecting their final course grade.

Learning outcome metrics. Two co-authors, who were not involved in the creation of AI-only group files, analyzed all anonymized notebooks from all three groups. Each notebook was evaluated along five main dimensions (Visual quality, Storytelling, Illustration Quality, Content relevance, and Notebook quality) on a scale from 1 (poor) to 5 (excellent) and combined into an average *Assignment score*. The measures relied on students adequately including visual design principles in their productions, as well as following standard storytelling approaches covered in the course [12]. The relevance of the visualization and illustration is evaluated based on how fitting it is to a leaflet showing the impact of a charitable organization. For instance, line charts showing evolution over the years are considered a suitable type of visualization. In contrast, pie-charts or bar-charts showing only the differences between measures in different countries in a single year are less relevant. Finally, notebook quality was assessed based on how effectively it was used to structure the assignment, such as organizing information clearly, grouping related content, and separating code from narrative text.

Two independent experts graded the notebooks. We measured inter-rater reliability on the total score for each notebook using the quadratic weighted Cohen κ and Pearson correlation. The results of a quadratic weighted Cohen's κ ($\kappa = 0.690$), and a Pearson correlation ($r = 0.827$, $p < 0.001$) showed acceptable agreement between experts.

Learning process metrics. For the learning process, we used the items of the NASA task load index (NASA TLX [10]): *Mental Demand*, which assesses how mentally demanding the students perceived the task to be, *Physical Demand*; which assesses how physically demanding the students perceived the task to be; *Temporal Demand*, which assesses how time-pressured or hurried the students felt during the task; *Perceived Performance*, which assesses how successful students believed they were in completing the task (this item is reverse in a *Failure* item to calculate the combined TLX score); *Effort*, which assesses how much effort students felt they needed to invest to complete the task; and *Frustration*, which assesses the degree to which students felt frustrated, stressed, or discouraged during the task. The items are combined to create an index from 0-100.

Usefulness metrics. To evaluate the attitudes towards the usefulness of the GenAI tool, we developed a 4-item TAM-based questionnaire [19] with items

that asked about: (1) How useful GenAI was for this task, (2) how GenAI helped improve the quality of work, (3) how effective GenAI was to reducing the time spent on the task, and (4) if the student would want to use GenAI in the future for similar tasks.

4 Results

A total of 69 Jupyter notebooks were included in the study, with 22 assigned to the H group, 21 to the HAI group, and 26 to the AI group. As illustrated in Figure 2, the two student groups were comparable in terms of age, with a mean age of 23.55 years ($SD = 5.57$) in the H group and 21.48 years ($SD = 2.93$) in the HAI group. Gender distribution was balanced across groups, with 36% female in the H group and 33% female in the HAI group. No significant differences were found between groups in demographics, i.e., Age ($t = 1.54$, $p = .135$) and Gender ($z = 0.21$, $p = .835$) or academic achievement, i.e, Grade ($t = 1.35$, $p = .186$).

4.1 Learning outcomes (H vs. HAI vs. AI)

We conducted a one-way analysis of variance (ANOVA) to compare the effects of three experimental conditions (H, HAI, and AI) on the Assignment Score. The results are visualized in boxplots with significance levels in Figure 3 for the Assignment Score as well as for each of its five dimensions. The results show a significant difference between the groups for the Assignment Score ($F = 10.466$, $p = .0001$). This significant difference is also found in three out of five dimensions of the score (Storytelling, Illustration quality, and Notebook quality). Visually, it seems that for the Assignment Score, as well as for its composing dimensions, the more automated the task (moving from all humans to all AI), the better the score. To confirm this intuition, we conducted a post hoc Tukey HSD test for the Assignment Score (see Table 1).

The results show that the AI group appears to be significantly superior to the other groups with very large effect sizes (Cohen's d > 1.0). Interestingly, the results do not detect any significant differences between the AI+H and the H groups. It should be noted that when conducting this post hoc analysis on the composing dimension of the Assignment Score, a similar picture arises, AI scores significantly better than both other groups on the Storytelling score, the notebook quality and the illustrative quality. However, except for the Illustration

Fig. 2. Comparison of baseline characteristics between the H (N=22) and HAI (N=21) groups with p values.

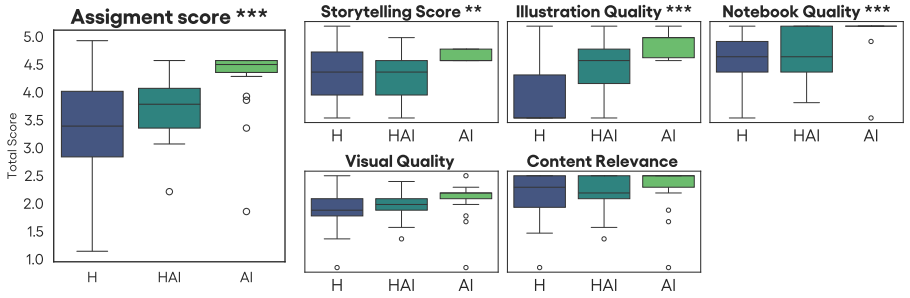

Fig. 3. Assignment scores and sub-dimension ratings for each group (H, HAI, AI). Left: overall Assignment Score; Right: breakdown across sub-dimensions. Sig. levels from ANOVA tests are indicated with asterisks: p $<$ .05 (*), $<$ .01 (**), $<$ .001 (***).

Quality, HAI performs better. The HAI and H groups are not statistically different in 4 out of 5 dimensions. As such, students with or without GenAI performed similarly on this task.

4.2 Learning process (H vs. HAI)

To understand if there were differences in how students perceived the task, we conducted a series of t-tests between the two groups of students on the task loading index (TLX) scores (see Figure 4). Overall, the TLX score was significantly higher in the group without GenAI ($M = 50.00$, $SD = 16.49$) compared to the group with GenAI assistance ($M = 31.47$, $SD = 16.91$), $t = 3.636$, $p = 0.0008$, with a very large effect size (Cohen's $d = 1.11$). As such, students who had access to GenAI found the task easier than those who did not. All dimensions of the TLX, except Physical load score, are significantly higher for students without access to GenAI. Interestingly, students with GenAI had a higher perception of their Performance (reversed Failure), even though this did not translate into actual improved performance as measured by their assignment scores.

4.3 Chatbot perception (HAI)

To understand if students were able to adequately judge whether GenAI was an asset for their task, we conducted a linear regression with the Assignment Score as the dependent variable and GenAI usefulness as the main predictor, using data from the *HAI group*. The model included Gender, Age, and Grade as control

Comparison	Mean(SD)	Mean(SD)	Mean Diff	95% CI	p-adj	Sig.	Cohen's d
AI vs. H	4.29 (0.57)	3.40 (0.90)	–0.89	[–1.37, –0.42]	<.001	***	1.18
AI vs. HAI	4.29 (0.57)	3.73 (0.54)	–0.56	[–1.04, –0.08]	.018	*	1.02
H vs. HAI	3.40 (0.90)	3.73 (0.54)	+0.33	[–0.17, +0.83]	.264		–

Table 1. Tukey HSD post-hoc comparisons for assignment scores between groups (H, HAI, AI). Table includes group means and standard deviations, mean differences with 95% confidence intervals, adjusted p-values, sig. levels, and Cohen's d effect sizes.

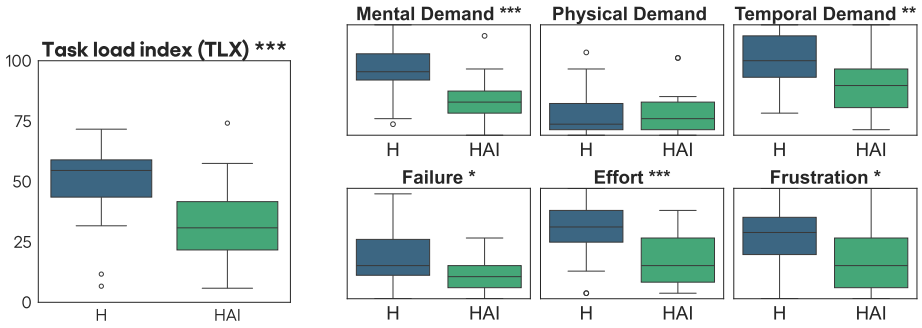

Fig. 4. Perceived workload ratings (NASA-TLX) for the H and HAI groups. Left: total score; Right: overview for each dimension of the NASA-TLX score. Significance levels from T-tests are indicated with asterisks: p $<$ .05 (*), $<$ .01 (**), $<$ .001 (***).

variables. The regression showed a significant *negative* effect of GenAI usefulness on the Assignment Score ($\beta = -3.23$, $p = .025$). Gender, Age, and Grade were not significant. Table 2 presents the statistics, while Figure 5 shows the results visually. The model explained 36.8% of the variance in Assignment Score ($R^2 = .368$; adjusted $R^2 = .210$; $F(4, 16) = 2.33$, $p = .100$). The standardized coefficient for GenAI usefulness was $\beta = -0.55$, and the semi-partial R^2 indicated that it uniquely explained 24% of the variance in the Assignment Score (a medium to large effect). A reason for this negative correlation could be that students who are less proficient academically might find GenAI more useful, and even if their results are worse, they could nevertheless have benefited from the help of GenAI. However, since we controlled for academic achievement by including the Grade variable as a control, this explanation seems less plausible. As such, these results seem to show that students were not only not able to see when GenAI is useful for them (an absence of correlation would have supported that claim), but they also seemed to be misled as to when GenAI is useful.

4.4 Qualitative insights into GenAI user experience (HAI)

Besides collecting quantitative data, the survey gathered qualitative feedback on user perceptions about their experience during the task.

Perceived benefits. Reflecting the quantitative results, participants in the HAI group perceive GenAI to be generally "easy to use" (P17, M, 21y), very helpful

Predictor	**β (raw)**	**β (std)**	**SE**	**p-value**	**95% CI**	**Semi-partial R^2**
Intercept	37.84	–	16.79	0.039	[2.25, 73.43]	–
GenAI usefulness	–3.23	–0.55	1.31	**0.025**	[–6.00, –0.45]	0.240
Gender	–2.58	–0.33	1.59	0.124	[–5.95, 0.79]	–
Age	0.35	0.27	0.42	0.416	[–0.53, 1.23]	–
Grade	–0.34	–0.06	1.72	0.846	[–3.99, 3.31]	–
Model fit: $R^2 = 0.368$, Adjusted $R^2 = 0.210$, $F(4, 16) = 2.33$, $p = 0.100$						

Table 2. Linear regression predicting Assignment Score (TOTAL) from GenAI usefulness, controlling for Gender, Age, and self-reported Grade. The table includes unstandardized and standardized coefficients, and semi-partial R^2 for the main predictor.

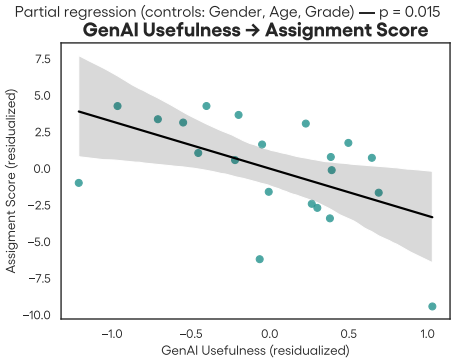

Fig. 5. Partial regression plot showing the relationship between *GenAI usefulness* and *Assignment Score*, controlling for Gender, Age, and Grade.

and efficient for their project. Descriptions such as *"very efficient tool, understands almost everything you want to do"* (P35, M, 19y) and *"The time required for the task is considerably reduced"* (P27, M, 25y) were common. Several participants also reported that AI assistance reduced their cognitive load. As one explained, *"ChatGPT has helped me remember certain details that I wouldn't have thought of. It also reminded me of things I'd forgotten. Overall, it decreased my mental workload in relation to the task in hand"* (P41, F, 27y). The use of AI was described as helpful and enhanced overall productivity. As participants stated, *"It's a tool that saves time and also allows you to learn in a more practical and effective way"* (P26, M, 28y), and *"ChatGPT makes the task look easy"* (P33, M, 19y). Several participants further emphasized the pedagogical and motivational impact of the tool. One referred praised GenAI for assisting them in *"finding solutions to the problems people face"* (P20, F, 21y). At the same time, another described it as *"a teacher who is always available, which is very motivating"* (P25, M, 21y). Another participant reported that the tool enabled a rapid transition from ideation to execution, stating that *"ChatGPT makes it quite easy to have a good brainstorm and then come up with a solid project base quickly, which allowed me to concentrate on developing the content"* (P30, M, 25y).

Perceived risks. In addition to the perceived benefits, some participants identified risks associated with the use of GenAI. Despite a participant explaining that *"the use is practical and allows non-completer or beginner students to succeed by understanding"* (P13, F, 20y), several others questioned how useful the technology is for beginners and how easy-to-use the solution really is. One participant noted challenges related to prompting, stating that it was *"very efficient, but the request made to ChatGPT must be precise and clear"* (P06, F, 24y). Another expressed concerns about the quality of GenAI's generated content, noting that *"the output was not qualitative"* (P37, M, 21y). Others questioned the tool's ability to understand users' requests. As P23 noted, *"Using ChatGPT was moderately helpful to me because it took a long time for it to understand exactly what I was asking. It was complicated"* (P23, F, 20y). P10 (M, 21y) also noted that *"it*

Category	Example Prompt(s)
Analysis (28.8%) Technical queries focused on programming help, troubleshooting, or debugging.	*"Correct this code: [Copy-paste error message]"; "Hi, could you generate some code for me to upload .csv content to python so that I can use the content to create visualizations?"*
Storytelling (17.4%) Requests for written content, ideas, explanations, or help making sense of text.	*"I have this database, and I need to find some interesting plots to do, which variables pair well?"; "Add emotion"; "'Percentage of the population with access to sanitation' What does this sentence mean?"*
Visual (11.5%) Design-related prompts, including aesthetic adjustments and visual consistency.	*"Perfect, is there any way the colours can match (each country has its own colour)?"; "For the first graph I want to thicken the lines"; "Generate an illustrative photograph that can be added to reinforce the impact of your visualisations"*
General (8.7%) General questions, sometimes minimal interaction (i.e., copy-pasted task instruction)	*[copy-paste the instructions], "Show it to me directly", "Can you do the same thing with my database?"*
Off-topic (31.3%) Prompts unrelated to the task sent before the experiment began. This category is included to consider the impact of students performing the task in an old conversation threat.	*[student starts the exercise on an old conversation]; "Explain to me the concept of frontiers of opportunity in economics"*
Other (2.4%) Simple interaction words	*"Perfect"; "Thank you"; "Continue"*

Table 3. Prompt categories, descriptions, and example prompts

took several attempts to get what I wanted. I think someone with the right skills on the program [without ChatGPT] might have been quicker". Furthermore, one participant highlighted risks related to hindering their critical thinking, stating that *"it [ChatGPT] prevented me from relying on my ideas, even though the AI was good. The AI ideas sometimes made me forget my critical mind"* (P16, M, 19y).

These insights suggest that GenAI is perceived as an easy-to-use tool that provides an immediate sense of usefulness. However, upon closer inspection, it appears that users need more time and skill than they initially expected to get exactly what they want from the tool.

4.5 Chatbot conversation analysis (HAI)

Participants in the HAI group were asked to share the conversations they had with ChatGPT while completing their assignments. We received data from 86% of the participants (18 out of 21). Students submitted 288 prompts in total ($M = 16.3$, $SD = 17.3$). Usage varied substantially, ranging from 3 to 72 prompts per participant. We analyzed the content of each prompt and coded it into one of six categories based on how participants used GenAI: General, Analysis, Visual, Storytelling, Off-topic, and Other (see Table 3 for details and example prompts).

To gain additional insights into how students who found GenAI to be more useful than average utilized the tool, we conducted case studies of the two highest performers and the two lowest performers based on the Assignment Score.

The first low performer (P26, M, 28y, score 3.4), who has the second highest number of total interactions (45), appears to have started the assignment in an

old conversation. In fact, 39 of them fell into the Off-topic category, all of which were generated before the assignment queries. Four of the remaining six prompts were used for analysis, mainly to help him with coding issues by copy-pasting the same error message three times in a row. This suggests that the student had difficulties correcting his code and struggled to get the answer he was looking for. The second low-performer (P10, M, 20y, score 2.2) used GenAI sparingly (9 prompts), despite finding it very useful. He also saw his own limitations, reporting that he needed several attempts to get results, and believed a skilled person might be faster without GenAI. This is reflected in his prompts, mostly related to code or debugging (5 out of 9), where he struggled to clearly express his needs. For example, after receiving several unsatisfactory responses, he ended up typing simple instructions like *"but say the code"*, reflecting a struggle to formulate effective queries.

The first high-performer (P6, F, 24y, score 4.5), understood that to be efficient, her instructions needed to be *"precise and clear"*. Her usage was minimal (4 prompts in total), consisting of two General prompts and two storytelling-related prompts. The prompts show a clear and targeted strategy: sending the assignment, uploading the necessary files, and asking for a French version of the output. This efficient use aligns with a low prompt count and a focused interaction. The second high performer (P20, F, 21y, score 4.5) took a different approach and performed iterative refinement. She performed an average number of prompts (16), mainly to improve her Storytelling (10 prompts) and visual design (4 prompts). The conversation reflects rich ideation and refinement, with repeated requests for additional alternatives such as *"another"* or *"many"*.

On one hand, these results confirm that GenAI's pontential can be unlocked by an adequate prompting strategy and by using the tool in a more creative and iterative way during exploratory processes, an aspect students seem to be aware of. On the other hand, the two low-performers repeatedly reformulated their prompts, suggesting that GenAI's potential is limited when students struggle to clearly express their needs.

5 Discussion

This research aimed to answer how GenAI tools can support learners in crafting data visualization stories (RQ). Our findings indicate that the introduction of GenAI into data visualization learning significantly decreased the perceived workload of students. As shown by the NASA-TLX results, students who used GenAI reported less mental and temporal demand to accomplish their tasks and expressed less frustration and fear of failure. The differences in these measures between groups are significant ($p<.001$) with a very large effect size (Cohen's $d=1.11$). These results complement the existing literature that highlights the potential of GenAI to scaffold complex tasks and reduce cognitive load in learning environments [13,46]. These findings, combined with the qualitative analysis of student comments, suggest that augmenting a data visualization task with GenAI can make it much less overwhelming for students and potentially more motivating to complete. Accordingly, GenAI seems to play a positive role in

lowering the entry barrier to data visualization learning by guiding users and reducing uncertainty, in line with the previous literature [5].

However, our results do not indicate an improvement in learning outcomes measured through Assignment Scores. The existing literature reported mixed results about the impact of GenAI on learning outcomes. While several studies highlight improvements [34,33], others observe declines [22,20], and some report no significant change [40]. In our context, GenAI integration did not lead to statistically significant differences in assignment performance. This suggests that while GenAI may have eased the learning process and reduced perceived workload, it did not necessarily translate into measurable gains in learning outcomes.

Based on our results, one could argue that GenAI's support improves the learning experience without reducing performance, which is a positive outcome. However, our investigation of the link between AI's usefulness and the Assignment Scores showed that there was a significant negative association between these two variables, controlling for Age, Gender and Grade. This finding seems to indicate that there was an illusion of usefulness rather than real learning support. This finding can be viewed in light of previous work that identified the risk of overreliance as a potential danger [15,3].

Furthermore, our results show that the AI-alone group failed to produce valid outputs in terms of format in 48% of the cases. But, when it did, it outperformed the groups involving students significantly. These findings suggest that the HAI group could have achieved better results simply by submitting the assignment instructions as a prompt and using the AI's output directly. However, students did not take this approach. This underlines the lack of skills in using these tools to their full extent, a concern highlighted in the previous literature [29,8]. This could potentially be mitigated by training students in using GenAI tools and improving their prompting skills, as highlighted in prior work (e.g., [30,20]). The conversations students had with GenAI show that low performers shared a pattern of engaging with ChatGPT in long, multi-topic conversations without starting a new chat for the study task. The earlier content in those threads likely introduced noise that may have impacted the model's ability to generate relevant and coherent responses. This aligns with known limitations of large language models, which can struggle with context retention and topic relevance in extended, multi-topic conversations [16]. In contrast, high performers used several efficient strategies, from a few precise prompts to repeated iterations to refine the output. These findings show that user prompting behavior and tool literacy emerge as central factors in determining the effectiveness of GenAI in educational tasks [18]. As such, it might be important to further integrate GenAI into a learning experience by constraining its capabilities and providing more task-focused interaction. As advocated by Mezzaro et al. [20], even if restricting access to the GPT model may limit students' freedom, it can support their learning. This approach could be leveraged for more open-ended projects such as the one discussed here.

Finally, this research is not without limitations. While a blind review process was used, reviewers were often able to guess which group an output belonged to,

which could have introduced bias in the evaluation. In addition, the relatively small sample size per group limited the ability to detect small effect sizes. Also, this study does not explore the long-term effects of GenAI integrations on learning outcomes; therfore, the future research is needed to understand sustained impact on students' skills and knowledge.

6 Conclusions

With the increasing need to teach students to craft convincing data stories, it is crucial to understand how GenAI can lower the entry barrier and lead to improved outcomes. To do so, we conducted a between-subject controlled experiment in a classroom setting (N=43) with a treatment group where students have access to GenAI (HAI) and a control group where they do not (H). Our findings show that when students have access to GenAI, the task load is significantly reduced with a very large effect size. However, the results show that the actual performance scores did not improve. What is more concerning is that the more students found GenAI useful, the worse their learning outcomes were, indicating a risk of overreliance. Finally, we also included a third group (AI alone), which outperformed the other two groups significantly in most performance qualities. This final result shows that GenAI offers promising possibilities for improved outcomes, yet learning experiences should be carefully scaffolded and students guided to take full advantage of GenAI capabilities.

Acknowledgments

This research was partially supported by the Swiss National Science Foundation (SNSF) under project number IZSEZ0_225217, and by Swissuniversities PgB project *STAGE: STEM Teaching in the Age of Generative AI*.

References

1. Bhaskaran, H., Kashyap, G., Mishra, H.: Teaching data journalism: A systematic review. Journalism Practice **18**(3), 722–743 (2024)
2. Clarizia, F., Colace, F., Lombardi, M., Pascale, F., Santaniello, D.: Chatbot: An education support system for student. In: CSS'18, vol. 10. Springer (2018)
3. DeJeu, E.B.: Using generative ai to facilitate data analysis and visualization: a case study of olympic athletes. Journal of Business and Technical Communication **38**(3), 225–241 (2024)
4. Denny, P., Prather, J., Becker, B.A., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B.N., Santos, E.A., Sarsa, S.: Computing education in the era of generative ai. CACM **67**(2), 56–67 (2024)
5. Ding, Z., Chan, J.: Intelligent canvas: Enabling design-like exploratory visual data analysis with generative ai through rapid prototyping, iteration and curation. arXiv preprint arXiv:2402.08812 (2024)
6. Firat, M.: What chatgpt means for universities: Perceptions of scholars and students. Journal of Applied Learning and Teaching **6**(1) (2023)

7. Fu, Y., Stasko, J.: More than data stories: Broadening the role of visualization in contemporary journalism. IEEE Transactions on Visualization and Computer Graphics (2023)
8. Garg, A., Rajendran, R.: The impact of structured prompt-driven generative ai on learning data analysis in engineering students. In: CSEDU'24. Science and Technology publications (2024)
9. Gottipati, S., Shim, K.J., Shankararaman, V.: Ai for connectivism learning - undergraduate students' experiences of chatgpt in advanced programming courses. In: AMCIS. AIS (2023)
10. Hart, S.G., Staveland, L.E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In: Advances in psychology, vol. 52, pp. 139–183. Elsevier (1988)
11. Hsin, W.J.: The effect of chatgpt: Student perspective and performance achievement. Journal of Computing Sciences in Colleges **39.0** (2024)
12. Kernbach, S.: Storytelling canvas: A visual framework for developing and delivering resonating stories. In: IV'18. pp. 390–395. IEEE (2018)
13. Khatib, R., Mattalo, B.: Enhancing learning experiences with genai chatbots: A tutorial approach. In: ICIS. AIS (2024)
14. Khemka, M., Houck, B.: Toward effective ai support for developers: A survey of desires and concerns. CACM **67**(11), 42–49 (2024)
15. Kim, N.W., Ko, H.K., Myers, G., Bach, B.: Chatgpt in data visualization education: A student perspective. In: VL/HCC'24. pp. 109–120. IEEE (2024)
16. Kim, Y., Lee, J., Kim, S., Park, J., Kim, J.: Understanding users' dissatisfaction with chatgpt responses: Types, resolving tactics, and the effect of knowledge level. In: IUI'24. p. 385–404. ACM (2024)
17. Liao, J., Zhong, L., Zhe, L., Xu, H., Liu, M., Xie, T.: Scaffolding computational thinking with chatgpt. IEEE Transactions on Learning Technologies **17.0** (2024)
18. Lyu, W., Wang, Y., Chung, T., Sun, Y., Zhang, Y.: Evaluating the effectiveness of llms in introductory computer science education: A semester-long field study. In: L@S'24. ACM (2024)
19. Masrom, M.: Technology acceptance model and e-learning. Technology **21**(24), 81 (2007)
20. Mezzaro, S., Gambi, A., Fraser, G.: An empirical study on how large language models impact software testing learning. In: EASE'24. ACM (2024)
21. Ochoa, X., Huang, X., Shao, Y.: Exploring the potential of generative ai to support non-experts in learning analytics practice. Journal of Learning Analytics (2025)
22. Ohm, M., Bungartz, C., Boes, F., Meier, M.: Assessing the impact of large language models on cybersecurity education: A study of chatgpt's influence on student performance. In: ARES '24. IEEE (2024)
23. Okonkwo, C.W., Ade-Ibijola, A.: Chatbots applications in education: A systematic review. Computers and Education **2**, 100033 (2021)
24. Ouaazki, A., Bergram, K., Farah, J.C., Gillet, D., Holzer, A.: Generative ai-enabled conversational interaction to support self-directed learning experiences in transversal computational thinking. In: CUI'24. ACM (2024)
25. Ouaazki, A., Bergram, K., Holzer, A.: Leveraging chatgpt to enhance computational thinking learning experiences. In: TALE'23. IEEE (2023)
26. Padiyath, A., Hou, X., Pang, A., Vargas, D.V., Gu, X., Nelson-Fromm, T., Wu, Z., Guzdial, M., Ericson, B.: Insights from social shaping theory: The appropriation of large language models in an undergraduate programming course. In: ICER'24. ACM (2024)

27. Peña-Fernández, S., Meso-Ayerdi, K., Larrondo-Ureta, A., Díaz-Noci, J.: Without journalists, there is no journalism: the social dimension of generative artificial intelligence in the media. El Profesional de la información **32**(2) (2023)
28. Peyton, K., Unnikrishnan, S.: A comparison of chatbot platforms with the state-of-the-art sentence bert for answering online student faqs. Results in Engineering **17**, 100856 (2023)
29. Qureshi, B.: Chatgpt in computer science curriculum assessment: An analysis of its successes and shortcomings. In: ICSLT '23: Proceedings of the 2023 9th International Conference on e-Society, e-Learning and e-Technologies. ACM (2023)
30. Rajala, J., Hukkanen, J., Hartikainen, M., Niemelä, P.: Call me kiran – chatgpt as a tutoring chatbot in a computer science course. In: Mindtrek '23 (2023)
31. Ranoliya, B.R., Raghuwanshi, N., Singh, S.: Chatbot for university related faqs. In: ICACCI'17. pp. 1525–1530. IEEE (2017)
32. Rodrigues, C., Reis, A., Pereira, R., Martins, P., Sousa, J., Pinto, T.: A review of conversational agents in education. In: TECH-EDU'22. Springer (2022)
33. Samarakoon, P., Asanka, D., Jayalal, S., Jayalath, N.: Analyzing the learning effectiveness of generative ai for software development for undergraduates in sri lanka. In: SCSE'24. vol. 7.0. IEEE (2024)
34. Sengewald, J., Wilz, M., Lackes, R.: Ai-assisted learning feedback: Should gen-ai feedback be restricted to improve learning success? a pilot study in a sql lecture. In: ECIS'24. AIS (2024)
35. Sethi, F.: Faq (frequently asked questions) chatbot for conversation. International Journal of Computer Sciences and Engineering **8**(10), 7–10 (2020)
36. Sheese, B., Liffiton, M., Savelka, J., Denny, P.: Patterns of student help-seeking when using a large language model-powered programming assistant. In: ACE '24. ACM (2024)
37. Sinha, S., Basak, S., Dey, Y., Mondal, A.: An educational chatbot for answering queries. In: IEM Graph'18, pp. 55–60. Springer (2020)
38. Su, J., Yang, W.: Unlocking the power of chatgpt: A framework for applying generative ai in education. ECNU Review of Education **6**(3), 355–366 (2023)
39. Sun, Y., Jang, E., Ma, F., Wang, T.: Generative ai in the wild: Prospects, challenges, and strategies. In: CHI'24. pp. 1–16 (2024)
40. Vadaparty, A., Zingaro, D., Smith, D.H., Padala, M., Alvarado, C., Benario, J.G., Porter, L.: Cs1-llm: Integrating llms into cs1 instruction. In: ITiCSE'24. ACM (2024)
41. Valverde-Rebaza, J., González, A., Navarro-Hinojosa, O., Noguez, J.: Advanced large language models and visualization tools for data analytics learning. In: Frontiers in education. vol. 9, p. 1418006. Frontiers Media SA (2024)
42. Welsh, M.: The end of programming. CACM **66**(1), 34–35 (2022)
43. Wenger, D., Hossain, M.S., Senseman, J.R.: Ai and the impact on journalism education. Journalism & Mass Communication Educator **80**(1), 97–114 (2025)
44. Yang, A.C.M., Lin, J.Y., Lin, C.Y., Ogata, H.: Enhancing python learning with pytutor: Efficacy of a chatgpt-based intelligent tutoring system in programming education. Computers and Education **7.0** (2024)
45. Ye, Y., Hao, J., Hou, Y., Wang, Z., Xiao, S., Luo, Y., Zeng, W.: Generative ai for visualization: State of the art and future directions. Visual Informatics (2024)
46. Yilmaz, R., Yilmaz, F.G.K.: The effect of generative artificial intelligence (ai)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. Computers and Education **4.0** (2023)