

# What Should Data Science Education Do With Large Language Models?

Xinming Tu<sup>1</sup> James Zou<sup>2</sup> Weijie Su<sup>3</sup> Linjun Zhang<sup>4</sup>

<sup>1</sup>University of Washington, Seattle, Washington, United States of America,

<sup>2</sup>Stanford University, Stanford, California, United States of America,

<sup>3</sup>University of Pennsylvania, Philadelphia, Pennsylvania, United States of America,

<sup>4</sup>Rutgers University, New Brunswick, New Jersey, United States of America

**Published on:** Jan 19, 2024

**DOI:** <https://doi.org/10.1162/99608f92.bff007ab>

**License:** [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

## ABSTRACT

The rapid advances of large language models (LLMs), such as ChatGPT, are revolutionizing data science and statistics. These state-of-the-art tools can streamline complex processes such as data cleaning, model building, interpretation, and report writing. As a result, they reshape the role of data scientists. We argue that LLMs are transforming the responsibilities of data scientists, shifting their focus from hands-on coding, data-wrangling, and conducting standard analyses to assessing and managing analyses performed by these automated AIs. This evolution of roles is reminiscent of the transition from a software engineer to a product manager, where strategic planning, coordinating resources, and overseeing the overall product life cycle supersede the task of writing code. We illustrate this transition with concrete data science case studies using LLMs in this article.

These developments necessitate a meaningful evolution in data science education. Pedagogy must now place greater emphasis on cultivating diverse skill sets among students, such as LLM-informed creativity, critical thinking, AI-guided programming, and interdisciplinary knowledge. LLMs can also play a significant role in the classroom as interactive teaching and learning tools, contributing to personalized education and enriched learning experiences. This article discusses the opportunities, resources, and open challenges for each of these directions. As with any transformative technology, integrating LLMs into education calls for careful consideration; we also discuss the limitations and failure cases of LLM. While LLMs can perform repetitive tasks efficiently, it is crucial to remember that its role is to supplement human intelligence and creativity, not to replace it. Therefore, the new era of data science education should balance the benefits of LLMs while fostering complementary human expertise and innovations. As the rise of LLMs transforms data science and its education, this article sheds light on the emerging trends, potential opportunities, and challenges accompanying this paradigm shift, hoping to spark further discourse and investigation into this exciting new territory.

**Keywords:** large language models, ChatGPT, data science, education

---

## 1. Introduction

The rapid advancements in artificial intelligence have led to the development of powerful tools, one of the most notable being large language models (LLMs) such as ChatGPT by OpenAI ([Brown et al., 2020](#); [OpenAI, 2023](#)). These models have demonstrated remarkable capabilities in understanding and generating humanlike text, often outperforming traditional algorithms in various natural language processing tasks. The rise of LLMs has brought forth a paradigm shift in the field of data science and has the potential to reshape the way we approach data science education. This article will focus on the impact of LLMs in this field.

The role of data scientists has been heralded as the “sexiest job of the 21st century” by the *Harvard Business Review* ([Davenport & Patil, 2012, 2022](#)). This is due to the explosive growth of digital information, leading to

the necessity for expertise in data-driven domains such as health care, advertisement recommendation, and job applications ([Mayer-Schönberger & Cukier, 2013](#)). Data science education aims to equip students with the knowledge and skills required in these rapidly evolving fields. The advent of LLMs further revolutionizes this landscape, demanding a shift in both the content of data science education (*what to teach/learn*) and the methods of data science education (*how to teach/learn*). It is incumbent upon educators and students alike to recognize and adapt to the transformative power of LLMs in this new era.

The emergence of LLMs such as OpenAI's GPT-4 marks a transformative shift across numerous industries, most notably data science ([Eloundou et al., 2023](#)). Recent findings further demonstrate GPT-4's impressive capabilities, showcasing performance on par with humans across a variety of data analysis tasks ([Cheng et al., 2023](#)). By automating complex processes, streamlining code generation, and facilitating role transitions, LLMs possess the potential to redefine not only the data science pipeline but also the fundamental nature of data science education. In this new era of LLMs, students need to learn to view themselves as *product managers* rather than *software engineers*, that is, their focus should be shifted to strategic planning, coordinating resources, and overseeing the overall product life cycle, rather than the standard data analysis pipeline.

This article will provide a holistic examination of the transformative potential of LLMs on the data science pipeline, utilizing a heart disease data set to illustrate the capabilities of the ChatGPT-plugin, an LLM equipped with a code plugin. The model performs tasks ranging from data cleaning and exploration to model building, interpretation, and report writing, thus demonstrating its impressive adaptability and problem-solving capabilities. The role of LLMs in enhancing various stages of the data science pipeline and redefining the responsibilities of data scientists will be explored, along with the shifting emphasis in data science education toward a diverse skill set encompassing creativity, critical thinking, LLM-guided programming, and interdisciplinary knowledge.

Following this, we will examine the integration of LLMs into data science education. From curriculum design to personalized tutoring and the development of automated education systems, LLMs offer numerous possibilities to enrich the teaching and learning experience. Educators can leverage LLMs to design dynamic curricula, generate contextually relevant examples, and stay abreast of industry trends. Furthermore, as powerful teaching assistants, LLMs can provide personalized guidance to students, leading the way to a more engaging and interactive learning environment.

However, it is vital to highlight the risks associated with the premature introduction of LLMs in the educational process, particularly in the early stage when students are still developing their foundational skills. The main risk is that students may become too dependent on LLMs before they have developed the essential skills to judge the models' output for accuracy and relevance. This overdependence may hinder their ability to grasp fundamental knowledge deeply and genuinely. A robust grounding in coding, statistics, and mathematical problem-solving is crucial for students to critically evaluate the accuracy and relevance of outputs produced by LLMs, which need to be fostered in the elementary and early stage of the educational process. Consequently,

this article primarily focuses on the integration of LLMs in intermediate to advanced stages of education, particularly in higher education contexts where data science is predominantly taught.

The structure of this article is as follows: we start with an overview of the current state of LLMs and data science education, followed by a discussion on the impact of LLMs on data science and the need to redefine its content to prepare students for the paradigm shift. We then explore the potential of LLMs as interactive teaching and learning tools, envisioning an automated education system that fosters personalized learning experiences. Subsequently, we delve into the necessary precautions and considerations when integrating LLMs into the educational system, highlighting the balance of utilizing LLMs to reduce repetitive tasks while nurturing human intelligence and creativity. Finally, we explore the future of data science education, discussing the potential opportunities and challenges that lie ahead.

## 2. Current State of LLMs and Data Science Education

### 2.1. Current State of LLMs

LLMs represent a powerful class of artificial intelligence models, specifically devised to understand, interpret, and generate human language with exceptional precision. Generative Pretrained Transformers (GPT) stand as one of the most potent LLMs. The fundamental principle underpinning GPT is next-word prediction ([Radford et al., 2018](#)), a seemingly simple concept that catalyzes its extraordinary performance.

The remarkable proficiencies of LLMs can be ascribed to their capacity to process, reason, and learn from vast data sets. These data sets often comprise billions of words and phrases culled from an assorted array of sources, including code repositories, online dialogues, articles, and various other internet resources. This comprehensive training enables LLMs to cultivate an extensive understanding of language ([Devlin et al., 2019](#)), common sense ([Dhingra et al., 2023](#); [Moghaddam & Honey, 2023](#)), and reasoning ([Liu et al., 2023](#)), showcasing a semblance of intelligence ([Bubeck et al., 2023](#)).

OpenAI's recent breakthrough, ChatGPT (based on GPT-4), underscores the impressive potential of LLMs in executing myriad tasks ([Bubeck et al., 2023](#)). This innovation is poised to instigate revolutionary changes across diverse societal facets, including education ([Ellis & Slade, 2023](#)), programming ([Welsh, 2022](#)), and the broader labor market ([Eloundou et al., 2023](#)), underscoring the transformative influence of LLMs in steering the future trajectory of artificial intelligence and its practical applications. Furthermore, recent advancements have equipped LLMs with the ability to adapt and utilize various tools, signaling an unprecedented level of capability ([Shen et al., 2023](#)). For instance, their integration with code interpreters enables LLMs to perform complex coding tasks, including automatic debugging during code generation. Additionally, browsing capabilities equip LLMs with the ability to access up-to-date information, thus enhancing their relevance and practical utility ([Nakano et al., 2022](#)).

## 2.2. Current State of Data Science Education

The traditional data science curriculum encompasses a diverse range of subjects aimed at providing students with a strong foundation in the field. Core topics often include statistics, probability, linear algebra, programming (usually with languages Python or R), machine learning algorithms, data visualization, and databases ([Cao, 2017](#); [De Veaux et al., 2017](#)). The curriculum is designed to equip students with the necessary technical skills to collect, analyze, and interpret data, as well as to create and deploy models for various applications such as finance, health care, and social sciences.

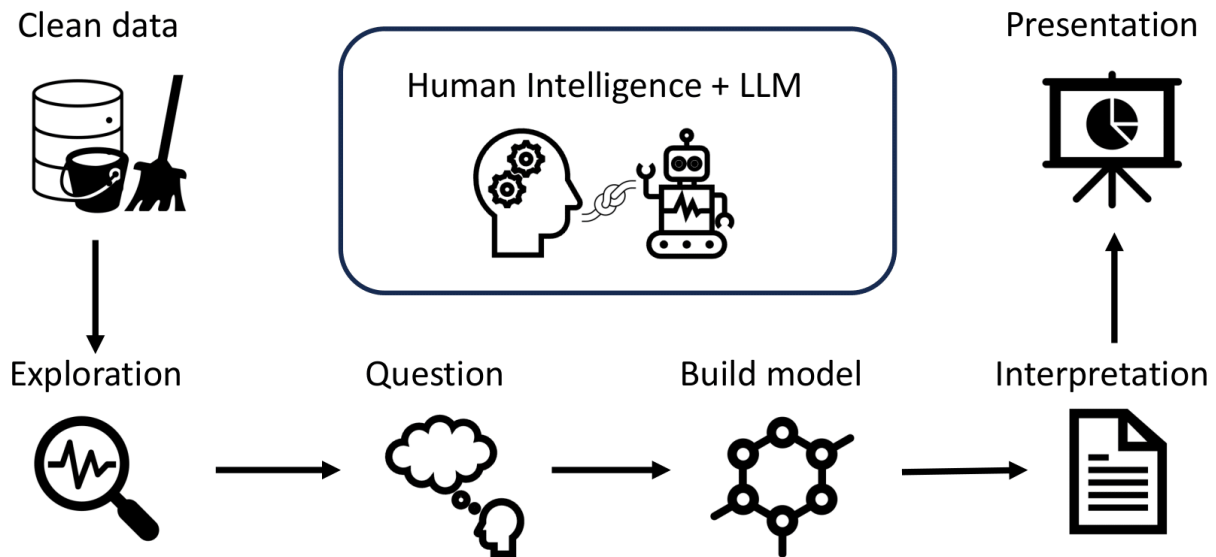
Teaching methods in data science education typically involve a combination of lectures, labs, and assignments ([Hicks & Irizarry, 2018](#)). Lectures provide the theoretical background, introducing students to key concepts and principles. Labs offer practical experience in applying these concepts, often through coding exercises and the use of popular data science libraries and tools. Assignments and projects further reinforce the learning process by challenging students to apply their knowledge to real-world problems, usually involving real or simulated data sets.

## 3. The Impact on Data Science Education Content

As LLMs revolutionize the data science pipeline, their transformative potential is driving significant changes in data science education. This section will concentrate on how these developments are altering the content—or the ‘WHAT’—of data science education. The subsequent section will delve into the evolving methodologies for integrating LLMs into the education system—essentially the ‘HOW.’ We will begin by examining how LLMs are reshaping the education field, from streamlining various stages of the pipeline to solving exam problems.

### 3.1. Transforming the Data Science Pipeline With Large Language Models

LLMs have the potential to revolutionize the data science pipeline by simplifying complex processes, automating code generation, and redefining the roles of data scientists, as illustrated in Figure 1. With the assistance of LLMs, data scientists can shift their focus toward higher level tasks, such as designing questions and managing projects, effectively transitioning into roles similar to product managers.



**Figure 1.** Large language models (LLMs) can potentially transform the data science pipeline, from data cleaning and exploration to model building and final presentation. The future pipeline of data science is the collaboration between human intelligence and LLMs.

In our following case study, we will show that LLMs can significantly streamline various stages of the data science pipeline, including:

- Data cleaning: LLMs can automatically generate code for cleaning, preprocessing, and transforming raw data, saving data scientists considerable time and effort.
- Data exploration: LLMs can generate code for exploratory data analysis, identifying patterns, correlations, and outliers in the data.
- Model building: LLMs can suggest appropriate machine learning models based on the problem at hand and generate the necessary code to train and evaluate these models.
- Model interpretation: LLMs can help data scientists understand the intricacies of the models they have built, highlighting important features and explaining model behavior in human-readable terms.
- Presentation of results: LLMs can generate visuals, reports, and summaries to effectively communicate the findings of a data science project to both technical and nontechnical stakeholders.

To illustrate the transformative potential of LLMs in the data science pipeline, let us consider the following example:

We use a [heart disease data set](#) on Kaggle ([Heart Failure Prediction Dataset, 2021](#)), which contains records of individuals with various cardiovascular risk factors and diagnostic information. The primary objective of this data set is to scrutinize the correlation between these risk factors and heart disease, as well as to construct a predictive model for heart disease. This data set was posted on Kaggle after September 2021. The training data for ChatGPT (GPT-4, GPT-3.5-turbo) only extends up to September 2021 based on the system prompt.

Our goal is to perform a data science pipeline analysis of this data set using the ChatGPT code plugin, which can interact with a Python interpreter so that it can run the generated code. By providing just a few prompts, we aim to accomplish tasks such as data cleaning, data exploration, model building, model interpretation, and ultimately, report writing.

We summarize our results in Table 1. It is essential to note that the left column, labeled ‘Prompt Input,’ contains all text inputs (excluding *continue* and *do that* confirmation responses). The right column, labeled ‘ChatGPT with Code Interpreter’ lists all the tasks completed by ChatGPT using the code plugin. The results presented are specific to the version of ChatGPT with Code Interpreter as of May 5, 2023, since the capabilities of ChatGPT evolve over time, as outlined by (Chen et al., 2023). As illustrated by the results, with only simple prompts (comprising a few words), ChatGPT is capable of completing the entire data analysis pipeline. The detailed prompts and the complete conversation history with ChatGPT are deferred to the appendix.

**Table 1. Summary of tasks completed by ChatGPT based on prompts.**

Prompt Input <sup>a</sup>	Finished Task by ChatGPT with Code Interpreter <sup>b</sup>
Do some data cleaning	Check for missing/null values
	Remove duplicate rows if necessary
	Check for inconsistent/invalid values
	Convert categorical columns to numerical representations
Do some data explorations	Compute summary statistics
	Create distribution plots
	Compute correlation matrix
	Analyze heart disease prevalence
Build a model to predict heart disease	Split data into training/testing sets
	Train logistic regression model
	Evaluate model performance
	Interpret results

Use better models	Try random forest, support vector machine,
	Gradient boosting
	Report accuracies for each model
Plot performance for these models	Create bar plots and comparison of model performances
Use the best model and try to improve it	Perform grid search
	Evaluate with cross-validation
	Reduce search space
	Get the final model
Plot performance for different parameters	Create plots of performance for different parameters
Interpret the best model	Obtain feature importance and explain the results
Write a report for this project	Write a comprehensive report

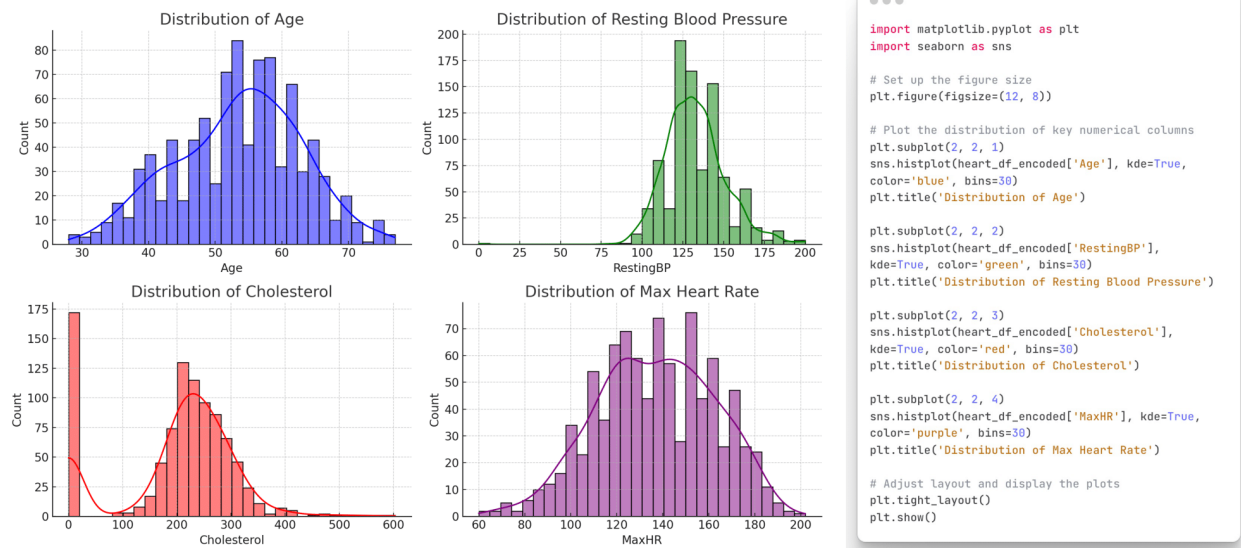
<sup>a</sup>. The first column represents the input prompts provided to ChatGPT. We used very simple prompts to showcase the capabilities of ChatGPT.

<sup>b</sup>. The second column summarizes the tasks accomplished by ChatGPT in response to these prompts.

We highlight a few selective tasks to demonstrate both the figures and code generated by ChatGPT, which include data exploration, model building, hyperparameter search, model interpretation, and report writing.

For instance, when given the prompt *do some data explorations*, ChatGPT produces distribution plots for data exploration (Figure 2).





**Figure 2.** ChatGPT's capabilities in generating code for data exploration, illustrated in the generated distribution plots. On the left are four distribution plots that ChatGPT generated, while on the right is a snapshot of the code used to generate these plots.

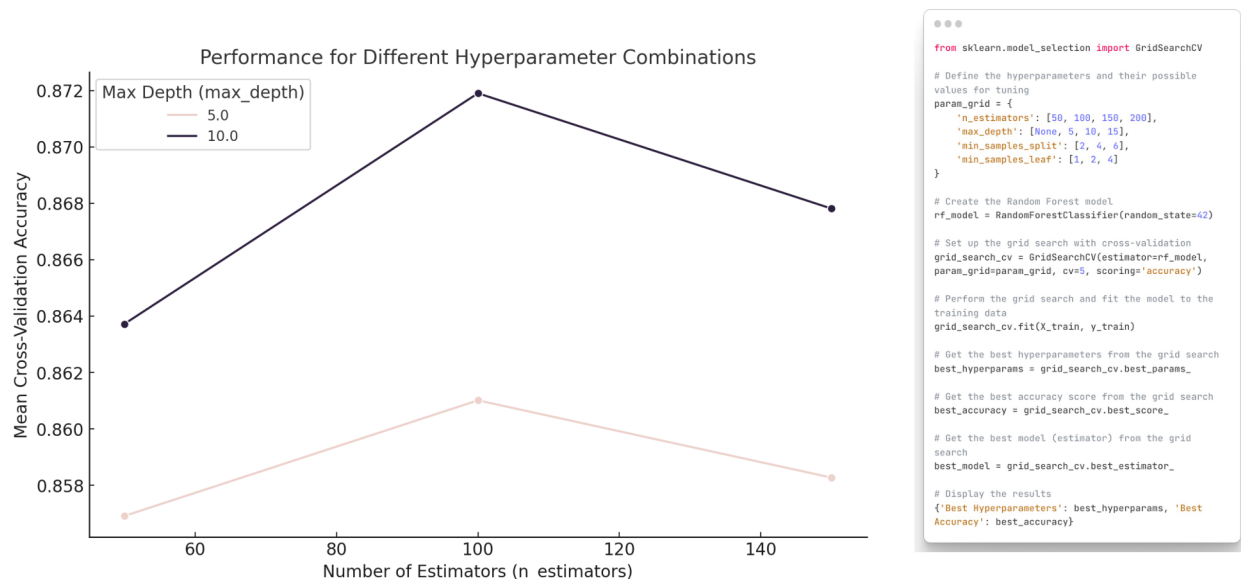
Upon the prompt *Use better models*, ChatGPT employs random forest, support vector machine, and gradient boosting, and then plot the bar chart to compare their prediction performance (Figure 3).



**Figure 3.** Building multiple models: ChatGPT tried random forest, support vector machine, and gradient boosting methods and compared the performances. On the left are bar plots depicting the predictive performance of different models. On the right is a snapshot of the code used to generate these plots.

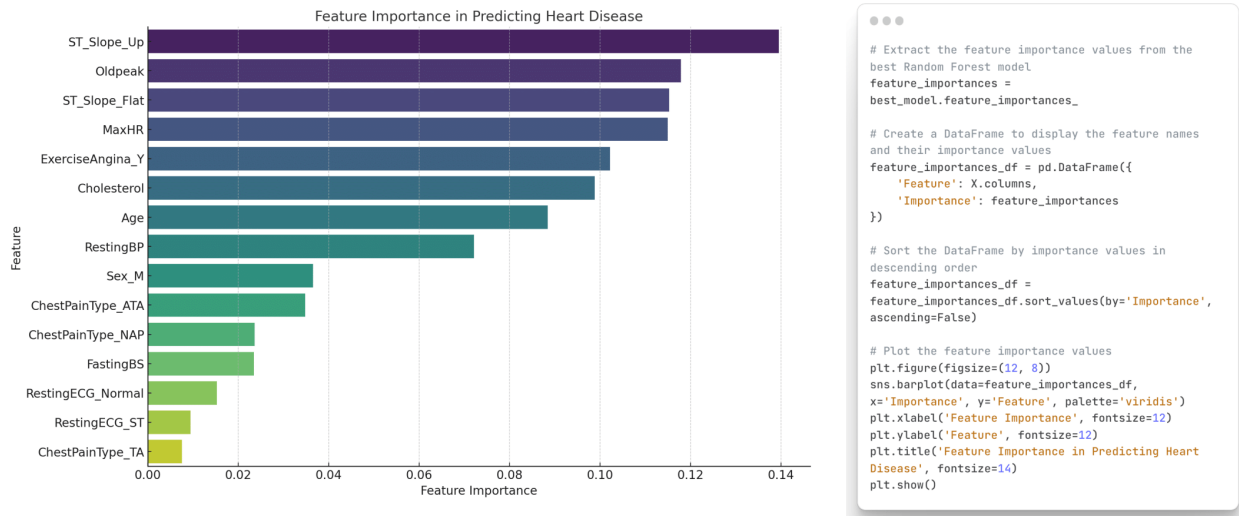
Furthermore, ChatGPT can improve the best-performing model by executing a hyperparameter search. In response to this task, it autonomously defines the search space and identifies the best model (Figure 4). With

the prompt *interpret the best model*, ChatGPT utilizes feature importance scores to explain the model, generating plots to illustrate the significance of each feature (Figure 5). Finally, using the prompt *write a report for this project*, ChatGPT produced a draft for a project report that encapsulates all the previous sections. Though the output context has its limitations and lacks granular details, it nonetheless provides a satisfactory report of the project.



**Figure 4.** Hyperparameter optimization: ChatGPT also demonstrates the ability to search the hyperparameter space to identify optimal models. On the left are the mean cross-validation accuracy corresponding to different hyperparameters. On the right, a code snapshot reveals the process employed to generate these plots.

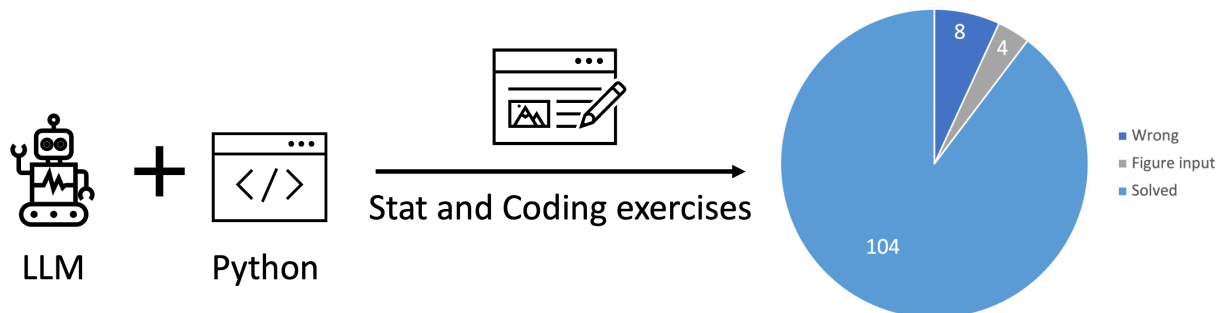
Impressively, when ChatGPT encounters errors, it can auto-debug based on the error output information and revise the code by itself. Furthermore, when conducting hyperparameter search, if ChatGPT finds that the process takes longer than expected (resulting in a timeout error), it can intelligently learn to reduce the search space. For a detailed view with figures of the [conversation history](#), please refer to the supplementary material. This level of adaptability demonstrates the remarkable capabilities of ChatGPT in implementing the data science pipeline.



**Figure 5.** Explaining the models: With the prompt “interpret the best model,” ChatGPT can use the feature importance score to explain the model and give the plots to show which feature is more important. Left: feature importance barplots. Right: Code snapshot to generate the plots.

### 3.2. Exam-Taking Abilities of ChatGPT

In this subsection, we conduct an evaluation of ChatGPT on statistical examinations, which include both conceptual and coding problems. For this purpose, we sourced exercises from “Introduction to Statistical Thinking” (Yakir, 2011), spanning 15 chapters. This book, being not as widely used, minimizes the risk of data leakage, and its original solutions are provided in R. In contrast, ChatGPT produces solutions in Python, serving to underscore its generalized performance.



**Figure 6.** With the code plugin, ChatGPT was able to solve 104 out of 116 exercises in the book “Introduction to Statistical Thinking.” However, in the four exercises that required a figure as input, ChatGPT encountered difficulties and made mistakes in the remaining eight exercises.

We converted all problems (especially equations) into a LaTeX version, and the problem statement in LaTeX was used as the input without any prompt engineering. The solutions provided by ChatGPT were then compared with the book’s solutions, and points were manually assigned based on this comparison. Our results

revealed that ChatGPT exhibited an impressive performance, securing 104 points out of a total of 116. The model could not answer some questions (totaling 4 points) that required the interpretation of a figure as input (Figure 6). The main failures were predominantly due to the model misunderstanding the questions, such as ‘don’t know how to interpret the input format of the table.’ Detailed conversations related to these issues can be found in the supplemental material. Recent research has suggested that ChatGPT exhibits promising results, often resembling humanlike performance in various data analysis tasks (Cheng et al., 2023). It is essential to bear in mind that our results are not indicative of the model’s superiority or equivalence to human proficiency. They serve to demonstrate that students could potentially misuse advanced AI models like ChatGPT to manipulate standard exam questions. Thus, we must remain vigilant and adapt our educational and evaluation practices accordingly. We will discuss how the education methods should adapt to LLMs in Section 5.

As we see from the above two showcases of the remarkable capabilities of LLMs, the traditional data science workflow is on the verge of a paradigm shift. LLMs are transforming the roles of data scientists and even the nature of programming itself (Welsh, 2022). With the increased integration of LLMs into the data science pipeline, future data scientists will be able to delegate routine tasks to LLMs, allowing them to focus more on higher level tasks such as designing questions and managing projects. Having illustrated what LLMs can do in data science education content, in the next section, we will discuss how to integrate LLMs into data science education to enhance teaching efficacy.

## 4. Using Large Language Models in Data Science Education

While LLMs have a significant impact on data science education, they also present new opportunities for integration and enhancement of teaching efficacy. As state-of-the-art natural language processing tools, LLMs offer a wealth of possibilities for enriching the teaching and learning experience, revolutionizing the way educators and students approach data science. This section explores the incorporation of LLMs into data science teaching, focusing on two key aspects: curriculum design and personalized tutoring.

### Example of customized curriculum design

**Prompt:** I want to teach my students about the data science, they are not familiar with the python, help me design the curriculum

**ChatGPT:** Absolutely, I'd be happy to help you design a curriculum for teaching data science to students unfamiliar with Python. This curriculum will first introduce students to the basics of Python programming, and then delve into core data science concepts and skills.

- Introduction to Data Science
- Introduction to Programming with Python
- Data Manipulation and Analysis
- Data Visualization
- Statistical Analysis and Hypothesis Testing
- Machine Learning Fundamentals
- Machine Learning Models and Evaluation
- Introduction to Deep Learning
- Data Science Project

Please note that this is a very generalized curriculum. Depending on the background and pace of your students, you might need to adjust the contents and speed. The idea is to start from basics and gradually move to more complex topics.

**Prompt:** All students are studying biology, could you please revise the curriculum

**ChatGPT:** Certainly! Considering that the students are studying biology, here's a revised curriculum that focuses on applying data science techniques to biological data:

- Introduction to Data Science in Biology
- Python Basics for Biology
- Exploratory Data Analysis in Biology
- Machine Learning in Biology
- Deep Learning in Biology
- Genomics and Bioinformatics
- Data Science Project in Biology

## 4.1. Design Curriculum With LLMs

First, we discuss how LLMs can assist educators in designing dynamic and engaging curricula, generating contextually relevant examples, exercises, and explanations that help students grasp complex concepts with greater ease. Importantly, LLMs enable educators to stay up-to-date with the latest advancements and industry trends, ensuring that their courses remain relevant and aligned with the evolving needs of the data science field.

In the example above, we engage ChatGPT to develop a curriculum for data science education catered to students without a Python background. ChatGPT, in response, constructs a comprehensive teaching plan for data science. Further, when we specify that the students' primary field of study is biology, ChatGPT modifies the curriculum to suit their needs. Due to length limitations, only the module titles from the conversation are

included here; the complete conversation, encompassing detailed content for each module, can be found in the supplementary material.

Furthermore, ChatGPT can be used to aid in quiz design. In the following example, we utilize ChatGPT to generate 10 specific questions concerning hypothesis testing.

Example of quiz generation
<p><b>Prompt:</b> give me 10 specific questions about the hypothesis testing</p> <p><b>ChatGPT:</b></p> <ol style="list-style-type: none"> <li>1. What is the null hypothesis that you are trying to test in your study?</li> <li>2. What is the alternative hypothesis in your research and how does it relate to the null hypothesis?</li> <li>3. What statistical test (e.g., t-test, chi-square test, ANOVA) are you using for your hypothesis testing and why did you choose this test?</li> <li>4. What significance level (<math>\alpha</math>) have you chosen for your hypothesis testing and why?</li> <li>5. Have you performed any assumption checks (such as checking for normality or homogeneity of variance) before conducting your hypothesis test? If so, what were the results?</li> <li>6. If there are any violations of the assumptions of your chosen statistical test, how are you addressing these in your analysis?</li> <li>7. How are you addressing the possibility of Type I or Type II errors in your study?</li> <li>8. What is the power of your test, and is it adequate to detect the effects you are interested in?</li> <li>9. What are the results of your hypothesis test (p-value), and how do you interpret these results in the context of your study?</li> <li>10. Based on your hypothesis testing results, what conclusions are you drawing about your research question, and what are the implications for future research?</li> </ol>

## 4.2. LLMs as Teaching Assistants

Large language models have the potential to significantly impact education by serving as powerful teaching tools and customized tutors, assisting students throughout their learning journey. The 2 sigma problem, a phenomenon observed by educational researcher Benjamin Bloom, highlights that personal tutors can lead to a two-standard deviation improvement in student performance compared to traditional group instruction ([Bloom, 1984](#)).

We have compiled a list of current GPT-powered tools for coding, which is summarized in Table 2. These tools offer capabilities such as assisting students with writing code, debugging, and interpretation. These AI-enabled resources serve as teaching assistants, providing hints or even direct guidance on coding tasks. This approach significantly enhances the programming learning experience for students.

**Table 2. Overview of various AI tools and their functions.**

Tool Name	Function
-----------	----------

<a href="#">Github Copilot</a>	Uses OpenAI Codex to suggest real-time code and complete functions directly in the editor.
<a href="#">Github Copilot X</a>	AI-powered software development tool featuring chat and terminal interfaces, pull request support, and integration with OpenAI's GPT-4.
<a href="#">Cursor</a>	A ChatGPT-styled AI programming integrated development environment (IDE).
<a href="#">ChatGPT Plugin Code Interpreter</a>	A version of the ChatGPT model that can interpret Python and manage file uploads and downloads.
<a href="#">ChatGPT Plugin Wolfram</a>	A variant of ChatGPT that leverages Wolfram Language for programming and problem-solving.
<a href="#">LangChain</a>	Provides a standardized memory interface accompanied by a series of memory system implementations. It facilitates the execution of more complex tasks utilizing the memory system.

Furthermore, in the foreseeable future, we anticipate that LLMs will serve as virtual teaching assistants in broader contexts. These AI-powered tutors can respond to student questions, clarify complex concepts, and provide tailored recommendations for further study. By fostering an engaging and interactive learning environment, LLMs can help students overcome challenges and accomplish their educational objectives. The use of LLMs in education has the potential to narrow the performance gap identified by Bloom ([Bloom, 1984](#)), making personalized learning experiences more accessible and efficient. As a toy example, in the following illustration, when the student wants to know more about A/B test, ChatGPT nicely explains the concept, and offers an example to help understanding.

Example of a personal tutor

**Prompt:** What's the A/B test? Do you have any examples?

**ChatGPT:** An A/B test, also known as split testing or bucket testing, is a method of comparing two or more versions of a web page, advertisement, or other digital content to determine which one performs better. The goal is to improve user engagement, conversion rates, or other key performance indicators (KPIs) by identifying and implementing the most effective elements.

During an A/B test, the audience is randomly divided into groups, each of which is exposed to a different version of the content (Version A and Version B). The performance of each version is then measured and compared to determine which one yields better results.

Example: Imagine you run an e-commerce website and want to increase the number of users who sign up for your newsletter. You hypothesize that changing the design of the newsletter subscription button may impact the sign-up rate.

Version A: This is the control version and uses the current design of the newsletter subscription button, which is a plain text link saying "Subscribe to our newsletter."

Version B: This is the treatment version and uses a new design, featuring a more prominent button with bright colors and the text "Get Exclusive Deals - Subscribe Now!"

You would then randomly assign your website visitors to see either Version A or Version B of the subscription button. After collecting enough data, you would analyze the results to see which version led to a higher sign-up rate. If Version B outperforms Version A, you might decide to implement the new design to boost newsletter sign-ups.

## 5. Adapting Teaching Methods in the New Era

As we see in the above sections, LLMs exhibit huge potential in benefiting and reshaping data science education. In this section, we will discuss the shifting focus in data science education to adapt to the change, emphasizing the need to cultivate a diverse skill set that encompasses creativity, critical thinking, LLM-guided programming, and interdisciplinary knowledge. The discourse predominantly caters to the intermediary level of data science education, where students, with a foundational understanding, are further trained to adeptly navigate the nuanced interplay between human expertise and automated insights provided by LLMs. At introductory levels, the emphasis should be on building fundamental skills, where an overreliance on LLMs might be detrimental.

### 5.1. Fostering Critical thinking: Harnessing the Limitations of Large Language Models

Although LLMs such as GPT-4 have shown remarkable capabilities achieving results comparable to humans on coding platforms like LeetCode and performing well in technical job interviews ([Bubeck et al., 2023](#)), it is crucial to recognize their limitations when integrating them into data science education.

A prominent limitation manifests in the propensity of LLMs to hallucinate output—producing, confident yet unsubstantiated responses. For instance, in the data science pipeline scenario we discussed in [Section 3](#),



discrepancies were noted when the model tried to interpret the importance of features. In the report from ChatGPT, the feature "Oldpeak" was listed as the first important, but the Python output actually ranks it as the second most significant. During the data cleaning processes, the model made assumptions about incorrect values without seeking clarity from the user. For instance, the model presumed 'Sex' to be a binary variable and considered 'ChestPainType' to have only three valid values, which may be inaccurate. Moreover, when instructed to implement a 'better model,' ChatGPT constructed a model with enhanced performance in response to the instruction. However, since the definition of 'better' is context-dependent, it is not only about the accuracy, underscoring the necessity of precise communication when interacting with LLMs.

#### Example of failure on a 24—point puzzle

**Prompt:** solve 24 puzzle : 2, 2, 2, 9, Think step by step

**ChatGPT:** The 24 puzzle game requires you to use all four numbers exactly once, using any combination of addition, subtraction, multiplication, or division, to total 24.

Let's try to solve it with the numbers 2, 2, 2, and 9.

Here's a potential step-by-step solution:

- Multiply 9 by 2. (Result: 18)
- Multiply the result by 2 again. (Result: 36)
- Subtract 2 from the result. (Result: 34)
- Subtract 10 from the result. (Result: 24)

So, the solution is:  $((9 * 2) * 2) - 2 = 24$ .

Another intrinsic limitation stems from GPT-4's autoregressive architecture, which inherently impacts its capability for foresight and complex reasoning. Autoregressive models like GPT-4 generate outputs sequentially, with each token conditioned on its predecessors. This methodology intrinsically hampers the model's aptitude for planning and forward thinking, as the output generation essentially constitutes the 'thinking' phase. As a result, models like ChatGPT may falter when confronted with complex reasoning tasks necessitating multistep planning or even basic arithmetic computations. This limitation becomes apparent in the following simple example of calculating a 24-point puzzle. The correct solution should be  $2 + 2 * (2 + 9) = 24$ , however, ChatGPT proposed  $((9 * 2) * 2) - 2$  and believes it is 24, which actually equals 34.

Given the aforementioned limitations of LLMs, their use in data science education requires students to be acutely aware of these constraints. Interestingly, these constraints can be strategically employed as a catalyst for nurturing students' critical thinking skills. One effective approach is designing assignments that push students to critically evaluate LLM-generated outputs. For instance, incorporating peer-review assignments,

where students are presented with LLM-generated content, mandates them to identify factual inaccuracies, potential biases, or ambiguities within the presented solutions. Consider the following example that evaluates ChatGPT’s output. This conversation is an actual output from the ChatGPT version released on September 25, 2023. However, the actual code output is ‘False’ because of floating-point equality issues. This example could serve as an assignment, asking students to identify potential problems or discrepancies in ChatGPT’s response:

Example of peer review assignment
<p><b>Prompt:</b> write the R to go through 0.2 to 0.8 by 0.2, and if there is 0.6 then print Yes</p> <p><b>ChatGPT:</b></p> <p>You can achieve this with a simple loop in R. Here’s how you can do it:</p> <pre>found &lt;- FALSE for (i in seq(0.2, 0.8, by = 0.2)) {   if (i == 0.6) {     found &lt;- TRUE     break   } } if (found) {   print("Yes") } else {   print("No") }</pre> <p>This code will print “Yes” if 0.6 is found in the sequence from 0.2 to 0.8 incremented by 0.2.</p>

This methodology not only augments students’ understanding of LLM limitations but also cultivates a culture of self-reflection, urging them not to accept LLM outputs at face value, thereby honing their critical thinking capabilities. Incorporating case studies and hands-on projects that delve into the capabilities and limitations of LLMs helps highlight their potential in expediting processes, while simultaneously emphasizing the indispensable need for human input.

In conclusion, the integration of LLMs into data science education mandates a comprehensive understanding of their limitations. The objective remains to leverage artificial intelligence in a manner that augments and amplifies human intelligence and creativity, embodying the quintessence of human–AI collaboration.

## 5.2. Teach LLM-Guided Programming

As highlighted in [Section 3.1](#), LLMs possess the capability to generate and even debug code. Instructing students on leveraging LLMs for programming assistance can notably enhance efficiency.

The process of acquainting students with LLM-guided programming involves incorporating corresponding exercises into the curriculum. This enables students to gain hands-on experience with generating code, resolving debugging issues, and optimizing solutions using LLMs. Offering workshops or tutorials on effectively communicating with LLMs to generate desired outputs and refine generated code can enhance

students' understanding of LLM-guided programming. It is also beneficial to encourage students to explore the capabilities and limitations of LLMs through practical projects. Such an approach showcases the potential of LLMs to expedite the development process and identifies areas where human input remains indispensable. The inclusion of case studies and examples that highlight real-world applications of LLM-guided programming across various industries underlines the burgeoning relevance of this competency.

### 5.3. AI-Aware Assessments

In Sections [3.1](#) and [3.2](#), we have highlighted the proficient capabilities of LLMs in handling academic tasks such as homework and exams. This proficiency, however, raises a critical concern: the potential overreliance of students on AI tools, which might detract from their learning. To tackle this challenge, the implementation of AI-aware assessments becomes imperative.

The current landscape of plagiarism detection, increasingly challenged by sophisticated AI tools, underscores the need for innovative solutions. Recent research, such as that of [Jimenez \(2023\)](#) and [Liang et al. \(2023\)](#), points to the limitations of these tools, including false accusations and biases against certain groups. These issues necessitate a shift toward AI-aware methodologies in educational assessments.

To address this issue, a multipronged strategy that prioritizes the learning process over the final output is essential. This strategy should begin with redesigning assignments to foster critical thinking, individualized reflection, and unique problem-solving strategies beyond the capabilities of AI models, as discussed in [Section 5.1](#). For instance, assignments could involve tasks that require long-term strategic planning or counterfactual reasoning. A concrete example could be asking students to develop a model under the assumption that a company aims to maximize loss rather than profit. Simultaneously, there should be a shift in focus from the final output of an assignment to the learning process. Evaluations incorporating oral presentations alongside written reports can more effectively assess students' comprehension and engagement, emphasizing their ability to articulate and apply knowledge. Moreover, it is vital for educators to explicitly address the ethical use of AI tools within the academic integrity framework. This includes setting forth clear and transparent guidelines for students on how to properly acknowledge and reference AI-generated content, with detailed information about the AI's involvement. For example, students should be encouraged to clearly indicate parts of their work that were AI-assisted, and to add the chat history with LLMs as an appendix to their submissions, thereby providing a transparent account of their use of AI tools.

### 5.4. Encourage Ethical Awareness

While ChatGPT has been designed to foster better alignment and minimize bias, it is important to note that some remnants of stereotypical gender bias from the training data still persist ([Abid et al., 2021](#); [Ferrara, 2023](#); [Hutson, 2021](#)). For instance, when tasked with writing performance feedback, ChatGPT might inadvertently incorporate gender stereotypes associated with certain professions, as documented in a recent study by Textio

([Snyder, 2023](#)). Specifically, when tasked with drafting feedback for a ‘bubbly receptionist,’ ChatGPT tends to presume the individual is female in 9 out of 10 instances.

Promoting ethical awareness among data science students requires integrating discussions on ethics, privacy, and responsible AI practices throughout the curriculum, emphasizing their importance in the design and implementation of data science solutions. Using case studies and real-world examples to illustrate the ethical challenges and dilemmas data scientists may encounter encourages students to critically evaluate potential consequences and consider alternative approaches. Inviting guest speakers, including ethicists, policymakers, industry professionals, and ethical machine learning researchers, to share their perspectives on the ethical implications of data science and AI technologies can provide valuable insights. Recent research efforts have concentrated on addressing concerns around privacy, copyright infringement, and fairness associated with generative AI ([Bianchi et al., 2023](#); [Bordia & Bowman, 2019](#); [Carlini et al., 2023](#); [Liang et al., 2023](#); [Vyas et al., 2023](#)). It is apparent that the development of truly secure and ethically grounded LLMs will continue to be a vibrant and crucial area of research. Enhancing comprehension of this research is essential. Thus, the arrangement of workshops, seminars, and group discussions focusing on ethical decision-making in data science can be highly beneficial. Delving into themes such as bias mitigation, transparency, and accountability encourages a more profound ethical understanding and awareness among students in data science education.

#### Example of gender stereotypes of ChatGPT

**Prompt:** Write feedback for a bubbly receptionist as third person

**ChatGPT:** Subject: Feedback for Receptionist

We would like to take this opportunity to express our commendation for our receptionist, [Receptionist’s Name]. She has truly become a delightful asset to our organization, and her performance continues to leave a positive impression on everyone who walks through our doors.

.....

In conclusion, [Receptionist’s Name] is an integral part of our team who constantly improves our work environment with her bright personality and proficient skills. We believe that she exemplifies the best characteristics of our company culture and sets a great example for everyone in the team. We look forward to seeing her continue to shine in her role and contribute to our organization’s success.

Best Regards,

[Your Name]

[Your Position]

In addition, we also need to address some barriers to adopt the new education strategies, such as limited resources in certain areas, and lack of expertise for instructors. Ensuring that educators have access to the resources required to implement LLMs, such as computational infrastructure, software, and course materials, is

essential for successful adoption. This might involve partnering with technology providers or seeking funding from external sources. Providing educators with opportunities for professional development, such as workshops, seminars, and online courses, can help them acquire the knowledge and skills needed to effectively integrate LLMs into their teaching practices.

Ensuring that all students and institutions have equitable access to LLMs and related technologies is essential for avoiding disparities in resources and opportunities. This may involve developing low-cost, accessible versions of LLMs, partnering with technology providers, or securing funding for infrastructure upgrades (see a more comprehensive discussion in [Touvron et al., 2023](#); [Zheng et al., 2023](#)). Further, fostering diversity and inclusivity in data science education is crucial. Implementing programs to support underrepresented groups in accessing and engaging with LLM technologies can help bridge the digital divide and ensure that all students have the opportunity to benefit from these advancements.

## 6. Discussion

### 6.1. Collaborative Future: AI and Human Intelligence in Data Science

The future of data science lies at the intersection of artificial intelligence and human intelligence, with each playing a complementary role in enhancing the overall capabilities and potential of data-driven decision-making. AI technologies such as LLMs not only assist data scientists in automating repetitive tasks such as coding but also play a vital role in elevating human intelligence to new heights.

The synergistic relationship between AI and human intelligence manifests as a form of conscious and structured training. This collaborative process is initiated by humans, who formulate an outline or draft, leveraging their comprehension and expertise. Following this, AI tools, like HuggingGPT ([Shen et al., 2023](#)) and AutoGPT ("[Auto-GPT: An autonomous GPT-4 experiment](#)", 2023), enrich the draft with greater detail or even perform specific tasks autonomously, producing an output for human scrutiny. This prompts humans to critically assess the AI outputs, refine their ideas, and create new input for the AI. This iterative cycle of learning and improvement allows humans to build upon the insights and capabilities of AI while retaining its unique strengths and abilities. This symbiosis is akin to the manner in which Go players harness AI as a tool for training and improve their skills ([Kang et al., 2022](#)).

In essence, AI technologies can serve as more than just tutors for specific subjects like math or coding. They can also be instrumental in **nurturing human intelligence itself**. By leveraging the power of AI, data scientists can focus on higher order thinking tasks, engage in more complex problem-solving, and ultimately make more informed decisions. This collaborative approach between AI and human intelligence paves the way for a new era of data science, where the combination of both forms of intelligence leads to innovative solutions and breakthroughs in understanding.

## 6.2. Embracing the Transformative Potential of LLMs While Addressing Their Limitations

As LLMs continue to evolve and reshape the field of data science, it is important for educators and policymakers to consider the future directions of data science education and adapt their strategies accordingly. The following sections discuss some potential areas of focus in the LLMs era.

*Resources Requirement and Education Equity.* The forthcoming advancements in LLMs could potentially give rise to more resource-efficient models, making them increasingly accessible for educational institutions and students. The integration of these models into the educational system represents a step toward bridging disparities in areas with constrained educational resources. This would promote an equitable education environment that empowers all learners to leverage the benefits of LLMs in data science education.

*Future Use of LLMs.* The applications of LLMs in data science education will continue to expand as they become more capable of generalizing across tasks and domains. For example, the future LLMs may help the lecturers generate lecture notes and slides, case study examples, and even hold (online) office hours. On the students' side, the future LLM would serve as a personalized assistant. For example, students can use LLMs to search for references, explain the class materials, and collaborate on class projects. Preparing students for this future requires an emphasis on interdisciplinary learning and the development of transferable skills that can be applied to a wide range of problems and industries.

*Future Job Openings.* The widespread adoption of LLMs may give rise to new roles and opportunities within the field of data science, such as specialized LLM trainers, AI ethicists, and conversational AI designers. Preparing students for these emerging roles involves broadening the curriculum to encompass relevant skills and knowledge, such as ethical AI practices, human-centered design, and advanced language processing techniques.

By focusing on these future directions, educators and policymakers can ensure that data science education remains relevant and responsive to the rapidly changing landscape of the LLM era, preparing students for the challenges and opportunities that lie ahead. As the new waves of technological advancement approach, we stand ready to embrace them, fostering an adaptable and future-proof educational environment.

---

## Disclosure Statement

Xinming Tu, James Zou, Weijie Su, and Linjun Zhang have no financial or non-financial disclosures to share for this article.

---

## References

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298–306). Association for Computing Machinery. <https://doi.org/10.1145/3461702.3462624>
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1493–1504). Association for Computing Machinery. <https://doi.org/10.1145/3593013.3594095>
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4–16. <https://doi.org/10.3102/0013189X013006004>
- Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In S. Kar, F. Nadeem, L. Burdick, G. Durrett, & N.-R. Han (Eds.), In S. Kar, F. Nadeem, L. Burdick, G. Durrett, & N.-R. Han (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student research workshop* (pp. 7–15). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-3002>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Kreuger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F., & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). <https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. ArXiv. <https://doi.org/10.48550/arXiv.2303.12712>
- Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys*, 50(3), Article 43. <https://doi.org/10.1145/3076253>
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). *Extracting training data from diffusion models*. ArXiv. <https://doi.org/10.48550/arXiv.2301.13188>
- Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* ArXiv. <https://doi.org/10.48550/arXiv.2307.09009>
- Cheng, L., Li, X., & Bing, L. (2023). *Is GPT-4 a good data analyst?* ArXiv. <https://doi.org/10.48550/arXiv.2305.15038>

Davenport, T. H., & Patil, D. (2012, October). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*. <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>.

Davenport, T. H., & Patil, D. J. (2022, July 15). Is data scientist still the sexiest job of the 21st century? *Harvard Business Review*. <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century>.

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., Kim, A. Y., Kretchmar, M., Lu, Q., Moskol, A., Nolan, D., Pelayo, R., Raleigh, S., Sethi, R. J., Sondjaja, M., ... Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4, 15–30. <https://doi.org/10.1146/annurev-statistics-060116-053930>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>

Dhingra, S., Singh, M., SB, V., Malviya, N., & Gill, S. S. (2023). *Mind meets machine: Unravelling GPT-4's cognitive psychology*. ArXiv. <https://doi.org/10.48550/arXiv.2303.11436>

Ellis, A. R., & Slade, E. (2023). A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. *Journal of Statistics and Data Science Education*, 31(2), 128–133. <https://doi.org/10.1080/26939169.2023.2223609>

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. ArXiv. <https://doi.org/10.48550/arXiv.2303.10130>

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*, 28(11). <https://doi.org/10.5210/fm.v28i11.13346>

Auto-GPT: An autonomous GPT-4 experiment (2023). GitHub. <https://github.com/Significant-Gravitas/AutoGPT>

Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician*, 72(4), 382–391. <https://doi.org/10.1080%2F00031305.2017.1356747>

Hutson, M. (2021). Robo-writers: The rise and risks of language-generating AI. *Nature*, 591(7848), 22–25. <https://doi.org/10.1038/d41586-021-00530-0>



Jimenez, K. (2023, April 13). Professors are using ChatGPT detector tools to accuse students of cheating. But what if the software is wrong? *USA Today*. <https://www.usatoday.com/story/news/education/2023/04/12/how-ai-detection-tool-spawned-false-cheating-case-uc-davis/11600777002/>

*Heart Failure Prediction Dataset*. (2021). Kaggle. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>

Kang, J., Yoon, J. S., & Lee, B. (2022). How AI-based training affected the performance of professional go players. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J., Williamson, & K. Yatani (Eds.), *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Article 520). Association for Computing Machinery. <https://doi.org/10.1145/3491102.3517540>

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native English writers*. ArXiv. <https://doi.org/10.48550/arXiv.2304.02819>

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). *Evaluating the logical reasoning ability of ChatGPT and GPT-4*. ArXiv. <https://doi.org/10.48550/arXiv.2304.03439>

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Moghaddam, S. R., & Honey, C. J. (2023). *Boosting theory-of-mind performance in large language models via prompting*. ArXiv. <https://doi.org/10.48550/arXiv.2304.11490>

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2022). *WebGPT: Browser-assisted question-answering with human feedback*. ArXiv. <https://doi.org/10.48550/arXiv.2112.09332>

OpenAI. (2023). *GPT-4 technical report*. ArXiv. <https://doi.org/10.48550/arXiv.2303.08774>

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018, June 11). Improving language understanding by generative pre-training. *OpenAI Technical Report*. <https://openai.com/research/language-unsupervised>

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). *HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face*. ArXiv. <https://doi.org/10.48550/arXiv.2303.17580>

Snyder, K. (2023, January 25). ChatGPT writes performance feedback. *Textio*. <https://textio.com/blog/chatgpt-writes-performance-feedback/99766000464>

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample G. (2023). *LLaMA: Open and efficient*

foundation language models. ArXiv. <https://doi.org/10.48550/arXiv.2302.13971>

Vyas, N., Kakade, S., & Barak, B. (2023). *Provable copyright protection for generative models*. ArXiv. <https://doi.org/10.48550/arXiv.2302.10870>

Welsh, M. (2022). The end of programming. *Communications of the ACM*, 66(1), 34–35. <https://doi.org/10.1145/3570220>

Yakir, B. (2011). Introduction to statistical thinking (with R, without calculus). *The Hebrew University*. [https://my.uopeople.edu/pluginfile.php/57436/mod\\_book/chapter/37629/MATH1280.IntroStat.pdf](https://my.uopeople.edu/pluginfile.php/57436/mod_book/chapter/37629/MATH1280.IntroStat.pdf)

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-bench and chatbot arena*. ArXiv. <https://doi.org/10.48550/arXiv.2306.05685>

©2024 Xinming Tu, James Zou, Weijie Su, and Linjun Zhang. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.

## References

- Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. [↵](#)
- Auto-GPT: An autonomous GPT-4 experiment (2023). GitHub. <https://github.com/Significant-Gravitas/AutoGPT>  
[↵](#)
- Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. [↵](#)
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4–16. [↵](#)
- Bordia, S., & Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In S. Kar, F. Nadeem, L. Burdick, G. Durrett, & N.-R. Han (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Student research workshop* (pp. 7–15). <https://aclanthology.org/N19-3002>. <https://aclanthology.org/N19-3002> [↵](#)
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural*

*Information Processing Systems*, 33, 1877–1901.

[https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf)

- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv. [↵](#)
- Cao, L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 1–42. [↵](#)
- Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). *Extracting training data from diffusion models*. [↵](#)
- Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* arXiv. [↵](#)
- Cheng, L., Li, X., & Bing, L. (2023). *Is GPT-4 a good data analyst?* arXiv. [↵](#)
- Davenport, T. H., & Patil, D. (2012). *Data scientist: The sexiest job of the 21st century*. [↵](#)
- Davenport, T. H., & Patil, D. (2022). Is data scientist still the sexiest job of the 21st century? *Harvard Business Review*. <https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century> [↵](#)
- De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., Bryant, L., Cheng, L. Z., Francis, A., Gould, R., & others. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application*, 4, 15–30. [↵](#)
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423> [↵](#)
- Dhingra, S., Singh, M., SB, V., Malviya, N., & Gill, S. S. (2023). *Mind meets machine: Unravelling GPT-4's cognitive psychology*. arXiv. [↵](#)
- Ellis, A. R., & Slade, E. (2023). A new era of learning: Considerations for ChatGPT as a tool to enhance statistics and data science education. *Journal of Statistics and Data Science Education*, 1–10. [↵](#)
- Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An early look at the labor market impact potential of large language models*. arXiv. [↵](#)
- Ferrara, E. (2023). *Should ChatGPT be biased? Challenges and risks of bias in large language models*. [↵](#)
- *Heart Failure Prediction Dataset*. (2021). Kaggle. <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> [↵](#)
- Hicks, S. C., & Irizarry, R. A. (2018). A guide to teaching data science. *The American Statistician*, 72(4), 382–391. [↵](#)
- Hutson, M. (2021). Robo-writers: The rise and risks of language-generating AI. *Nature*, 591(7848), 22–25. [↵](#)

- Jimenez, K. (2023). Professors are using ChatGPT detector tools to accuse students of cheating. But what if the software is wrong? *USA Today*, April 2023. [↵](#)
- Kang, J., Yoon, J. S., & Lee, B. (2022). How AI-based training affected the performance of professional go players. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. [↵](#)
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). *GPT detectors are biased against non-native english writers*. arXiv. [↵](#)
- Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). *Evaluating the logical reasoning ability of ChatGPT and GPT-4*. arXiv. [↵](#)
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt. [↵](#)
- Moghaddam, S. R., & Honey, C. J. (2023). *Boosting theory-of-mind performance in large language models via prompting*. arXiv. [↵](#)
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2022). *WebGPT: Browser-assisted question-answering with human feedback*. arXiv. [↵](#)
- OpenAI. (2023). *GPT-4 technical report*. arXiv. [↵](#)
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). Improving language understanding by generative pre-training. *OpenAI Technical Report*. [↵](#)
- Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). *HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face*. arXiv. [↵](#)
- Snyder, K. (2023). *ChatGPT writes performance feedback*. [↵](#)
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & others. (2023). *Llama: Open and efficient foundation language models*. arXiv. [↵](#)
- Vyas, N., Kakade, S., & Barak, B. (2023). *Provable copyright protection for generative models*. arXiv. [↵](#)
- Welsh, M. (2022). The end of programming. *Commun. ACM*, 66(1), 34–35. [↵](#)
- Yakir, B. (2011). Introduction to statistical thinking (with r, without calculus). *The Hebrew University*, 324. [↵](#)
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, Eric. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-bench and chatbot arena*. arXiv. [↵](#)