Sprint: Structured Approach to Data Analytics

Introduction to AI and Statistics

# **Artificial Intelligence Introduction**

*(Approximate time to complete: 3-4 hours)*

It might seem that recently, almost everyone has been talking about AI. Even though the field itself has been around for many decades and the most popular applications have taken many shapes and forms, a particular AI technology has emerged recently that has captivated everyone’s attention - Large Language Models (also called LLMs). The most popular LLM tool is undoubtedly ChatGPT, which you have probably already heard of or even tried using. Using LLMs, it’s easy to note their incredible ability to mimic human-like responses and even display what some would describe as creativity. It is these characteristics of LLMs that make them so unique and exciting - the closeness to human interaction and the ability to perform creatively in new, unseen tasks.

In this part, you will first go through a video that will start by explaining the very basics of what LLM AI is. It will also delve deeper into additional topics, such as how to create good prompts (also called “prompt engineering”). As it is a video created by Google, you will notice that they will be introducing the tools offered by Google Cloud – which are actually quite powerful and easy to try out! We will also list additional LLMs that you can try though – there is no specific one that is necessarily always better than the others. For example, ChatGPT, even though it’s the most popular one, requires a paid subscription to access its most advanced version that uses GPT-4 (newest version of the model). What surprisingly few people know though is that Microsoft’s Copilot AI chatbot uses the same GPT-4 and is available for free!

Afterward, you will find a couple of suggestions on how to use LLMs in more creative and useful ways. Furthermore, we will also talk about the ways how AI can be misused, so that you know what to be careful about.

***Note:*** *while LLMs are technically just a small subset of what has traditionally been called “AI”, in current contexts, both are usually used interchangeably. We will often do the same in this course.*

***Note:*** *Keep in mind that the field of AI is advancing quickly, so changes are frequent. We recommend staying alert to potential updates in both tools and materials.*

* [Introduction to Large Language Models](https://www.youtube.com/watch?v=zBj4eWZFBzs)
* [Another explanation of what AI and LLMs are](https://www.youtube.com/watch?v=D2JY38VShxI)
* [A document you can refer to for prompt engineering](https://developers.google.com/machine-learning/resources/prompt-eng)
* (optional) [More technical introduction](https://www.youtube.com/watch?v=N3sMj-sBcLg)
  + in case you are more curious about the technical and mathematical side of LLMs

### **Pick AI tools to practice and immerse yourself**

* [ChatGPT](https://chat.openai.com/)
  + For the most advanced version (which adds image generation, customGPTs, GPT-4 model and more), a monthly subscription will be needed. One thing that is useful and does not require a subscription though is “custom instructions”. They allow you to make sure the AI always responds in a particular way, without forgetting it. To enable it, just click on your profile on the bottom left corner of the screen and select “Custom Instructions”.
  + (optional) For a more detailed presentation of exciting new developments in ChatGPT, see [Sam Altman's presentation on ChatGPT by OpenAI.](https://www.youtube.com/watch?v=U9mJuUkhUzk)
* [Gemini](https://gemini.google.com/)
* [Microsoft Copilot AI](https://www.bing.com/search?q=Bing+AI&showconv=1&FORM=hpcodx)
  + Formally Bing Chat, which is running on GPT-4 and is free
  + Can generate images for free. Try a prompt “Generate an image of a duck in space that is learning data analytics”
  + When it was released, it was the first GPT-4 model to the public (even though you had to join waitlists). Many argue, however, that Microsoft moved too quickly trying to jump on the “AI hype bandwagon” and released a product that was far from ready for public use. You can read a very interesting article about it [here](https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter) (full story requires a New York Times subscription).
* [Midjourney](https://www.midjourney.com/) is one of the most popular AI image generation tools. Initially, it gave new users free credits to generate several images, but now a subscription is required.

### **Direction for further research**

* LLMs can generate and run code. If you know some coding, or even want to learn the basics of it, try to ask an LLM to give you an example of a simple program. You can then ask how it works, even make some changes yourself and ask the LLM to run it. It will be able to explain the syntax and any errors that you make. If you are not sure which programming language to use, we recommend Python, as it uses a very easy-to-understand syntax.

### **Exercises**

* When studying over the next few days, try to look for opportunities to use LLMs to deepen your learning. Note down a particular case where an LLM surprised you and was particularly useful. Share this in the next stand-up that you join.
* Try to find a prompt that the LLM of your choice fails miserably at. Share this in the next stand-up that you join as well. There are quite a few areas that LLMs are really bad at (yet they sound very confident!).

## **AI study buddy**

While self-studying, consider using AI tools as your study buddy. They can assist you in assessing your knowledge. Create a new chat and start with a prompt related to the topic you've just studied for example, such as 'Google Sheet: Functions' Use this prompt as a reference for future sprints, modifying the topic or enriching the prompt with additional details to track your progress.

### **AI study buddy prompt**

\*I would like you to act as an interviewer on the topic 'Google Sheet: Functions'. I will assume the role of the candidate, and you will pose questions about the topic from various perspectives. Please respond only in the capacity of an interviewer. Avoid writing the entire conversation at once. The interview should be conducted exclusively with me. Pose the questions one at a time, as an interviewer typically would, and wait for my responses. Provide explanations only after I have submitted my answers. My first sentence is 'Hi'.\*

*Note: in ChatGPT, you can add this prompt as a custom instruction. This will ensure that the model always remembers to act as an interviewer.*

### **Dangers of AI**

1. Accuracy of Information: AI tools are designed to provide accurate and relevant information based on their training data and algorithms. However, the outcomes may not always reflect the most current or specific data available. Users are advised to cross-check critical information with reliable sources. A particularly dangerous form of inaccuracies in LLMs is called “hallucinations” – this happens when a model creates fake information and confidently gives that as a fact.
2. Data Privacy: When using AI tools, especially those requiring data input, be mindful of the information you share. Avoid submitting sensitive personal or confidential data that could compromise your privacy or security.
3. Dependence on AI: While AI tools are helpful, relying solely on them for learning or decision-making can limit your critical thinking and analytical skills. Use AI as a supplement to your research and knowledge.
4. Contextual Misinterpretation: AI may sometimes misinterpret the context or nuances of a query, leading to irrelevant or incorrect responses. It’s important to critically assess AI-generated content for its applicability to your specific needs.
5. Ethical Use: Employ AI tools ethically, avoiding manipulations that could result in harm, misinformation, or unethical outcomes. Respect the guidelines set by the tool’s creators.
6. Updates and Changes: AI tools evolve and improve over time. Stay informed about updates or changes in the tool’s functionality or data handling practices.
7. Limitations in Scope: AI tools may not cover every possible scenario or topic comprehensively. Their advice or output should be seen as one of several perspectives rather than an all-encompassing solution.

* [Responsible AI](https://www.youtube.com/watch?v=2Pzk7ySufZM)
* [A fun game where you need to trick the AI into revealing secrets](https://gandalf.lakera.ai/). Note: passing all the levels is extremely difficult. Passing 2-4 levels is more than enough to get an initial understanding of security issues in AI.

### **Using AI responsibly when learning**

* For guidelines on responsible and tactical use of AI tools, especially when it comes to learning at Turing College, please refer to the information provided in [this link](https://turingcollege.atlassian.net/wiki/spaces/DLG/pages/872480773/AI+tools+e.g.+ChatGPT+in+learning).

# **Additional Resource**

* (optional) [Interesting Experimental Analysis of the Tipping Prompt (0 to 1 Million $)](https://blog.finxter.com/impact-of-monetary-incentives-on-the-performance-of-gpt-4-turbo-an-experimental-analysis/)
* (optional) [Emotional Prompts](https://arxiv.org/pdf/2307.11760.pdf)

# **Time for Statistics**

Statistics plays a crucial role in the field of data analytics. It's not enough to just collect data; understanding it is key. Statistics allows us to turn complex data into clear insights by identifying patterns, predicting future trends, validating ideas with evidence and summarizing / describing data. This approach ensures our decisions are grounded in reality, not just guesswork, making our conclusions both trustworthy and actionable.

## **Required Material**

* [Introduction to Statistics](https://youtu.be/VPZD_aij8H0?si=yQfA-oWqRI9ryJhT&t=799)
  + Watch the video from the timestamp 13:19 (Why should you study statistics) to 27:50 (Randomness)
* [Descriptive vs Inferential Statistics](https://www.youtube.com/watch?v=VHYOuWu9jQI)
  + [Types of Statistics](https://www.geeksforgeeks.org/difference-between-descriptive-and-inferential-statistics/)
* (optional) [Data Reasoning in a Digital World](https://callingbullshit.org/syllabus.html)
  + If you find this interesting, go through all the modules bit by bit over the upcoming months

This segment was just the beginning, offering a glimpse into the critical role of statistics in data analytics. As we progress through subsequent sprints, we'll explore more specific and related topics in detail.

# **QUIZ**

1.What is an example of a descriptive statistic?

A.p-value

B.Mean

C.Confidence interval

D.Regression coefficient

2.What does “hallucinating” mean when it comes to AI?

A.An AI fabricating information that is false and providing it as fact

B.The ability for an AI to be creative

C.Visual content generation

D.None of the above

3.Which of these best describes Analytical thinking?

A.Using facts to guide business strategy.

B.The management of the people, processes, and tools used in data analysis.

C.The various elements that interact with one another in order to produce, manage, store, organize, analyze, and share data.

D.The process of identifying and defining a problem, then solving it by using data in an organized, step-by-step manner.

4.What is an important part of the Technical mindset?

A.The ability to break things down into smaller steps or pieces and work with them in an orderly and logical way.

B.The process of identifying and defining a problem, then solving it by using data in an organized, step-by-step manner.

C.The science of data

D.The various elements that interact with one another in order to produce, manage, store, organize, analyze, and share data.

5.Fill in the blank: In data analytics, fairness means ensuring that your analysis does not create or reinforce bias. This requires using processes and systems that are fair and \_\_\_\_\_.

A.Inclusive

B.Restrictive

C.Partial

D.Favorable

6.If you calculate the standard deviation of a set of exam scores, which type of statistics are you using?

A.Descriptive statistics

B.Inferential statistics

C.Both descriptive and inferential statistics

D.Neither descriptive nor inferential statistics