

Use of ChatGPT as a complement to data analysis and processing in RStudio

Almost daily we see news, social media posts, reports made by technology companies about artificial intelligence (AI) tools and, since 2022, when ChatGPT was launched, much of the attention has been focused on this particular tool.

Most of this information shows how AI tools available to end users have great applications and advantages in different areas, and that their level of sophistication and functionalities are improving very rapidly.

The field of physical data analysis is not outside of this evolution, and we can use tools such as ChatGPT as a complement in the processes that we have been discussing throughout the course.

ChatGPT (Chat Generative Pre-trained Transformer) is an artificial intelligence system based on a large language model (LLM). These models are trained from a large amount of text-type data so that they are able to interpret the user's language and generate responses as if it were a conversation between people. This type of LLM is based on deep neural networks, which learn from patterns in the text and its meaning (Almarie et al., 2023). Therefore, ChatGPT is an interface that allows interaction with the created model, in which the user can ask questions and the interface will return the answers as if it were part of a conversation.

We have mentioned certain concepts that may be new to our field; and we will use ChatGPT to clarify them.

Figure 1: Definition of Artificial Intelligence, Large Language Model and ChatGPT



Puedes definir los conceptos Inteligencia artificial, large language model y chat-gpt de manera sencilla?



Claro, aquí tienes definiciones sencillas de esos conceptos:

1. **Inteligencia Artificial (IA):** Es la capacidad de las máquinas de realizar tareas que normalmente requieren inteligencia humana, como el aprendizaje, la resolución de problemas y la toma de decisiones.
2. **Large Language Model:** Es un tipo de modelo de inteligencia artificial diseñado para comprender y generar texto en lenguaje natural con una gran cantidad de datos y capacidades de aprendizaje automático avanzadas.
3. **ChatGPT:** Es un ejemplo de large language model desarrollado por OpenAI. Está diseñado para mantener conversaciones naturales con los usuarios, responder preguntas, ofrecer asistencia y generar texto coherente y relevante en función del contexto proporcionado.

Source: screenshot taken from OpenAI. <https://openai.com/chatgpt/>

The main applications of this tool are to provide definitions, summarize content, etc. But we should note that it has certain limitations too. Like any machine learning or artificial intelligence model, it is trained with specific data; therefore, the characteristics of the data that have been provided to the model will condition the answers. This has clear implications on the output we receive, for example, in the first free and available versions of ChatGPT, it was reported that the model had been trained with data up to a specific date, which meant it would not be a useful tool to answer questions about current or recent affairs, since the model did not have that information. This tool is not a search engine like *Google*, so it will not provide links or landing pages to other resources.

To get the best possible output, as it is a language-based model, one needs to be careful in the way to request information from ChatGPT. The way we structure our question or request will condition the detail and accuracy of the answer. The questions we ask ChatGPT are known as "prompts". For this reason, we should follow a series of guidelines developed below.

- Include context before the question:
 - Who the supposed person asking the question is (a student, a coach, etc.) and to whom the answer is addressed;
 - What the objective of the question is.
- What information you want to get:



- be as specific as possible;
- Specify whether there is certain information that is not necessary
- Specify the length of the response.
- Ask for examples.

Example of a "bad" prompt:

- 3D graphic RStudio tennis serve.

Example of a "good" prompt:

- Could you describe the process of creating a 3D graphic in RStudio to visualize the trajectory of a tennis player's arm during serve? I'm interested in learning about different possibilities, useful libraries, and code as an example.

One of the particularities of ChatGPT is that we can track the questions we ask, that is, if we want it to extend its answer at any of the points we can ask it to expand on that specific point, to use other examples, etc.

These points only describe the format of the questions should be, but the most important point, as we have highlighted multiple times during the course, is the choice of the question itself. First, we need to define the specific concern in our context for which we seek an answer and then determine how ChatGPT can help us during the process.

The link to access the tool is:

OpenAI. (2021). ChatGPT. <https://www.openai.com/chatgpt/>

The use of ChatGPT during the data analysis process

In the video material, we will see examples on the use of the tool to perform queries that may arise on our physical data analysis projects. In this course, many types of analysis and functionalities of RStudio have been described to try to present its full potential. RStudio is a tool with a high learning curve, since, at the beginning, it requires a correct understanding of the structure that can be complex if it is the first programming language we work with. ChatGPT can help us solve doubts during the process, as we will see in the following examples.

- Understanding code: we can ask ChatGPT to describe what each of the parts of the code used are and what objective they achieve/pursue.
- Writing code: we may need help removing missing values, adding filters to columns, importing data, etc., before continuing with the analysis processes. To help, we can describe them to ChatGPT so that it can propose the specific code to meet those needs.



Please note this content has been included in the last course, since in the learning process of a tool like RStudio, it is important to understand from scratch what is achieved with each line of code, experiment with the different functions and look for solutions on our own to transform or modify the data that we create, as well as understand where the errors we make throughout the development of the code come from and how to correct them. If we use the information provided by ChatGPT without understanding what is happening, we can miss errors that condition the final results and limit our ability to learn the tool itself, which will condition its more extensive development in the future.

Once we have more experience with RStudio, ChatGPT is still a good complement to the projects we work on. As always, our goal is to automate our day-to-day processes of the Sport Scientist, so that the information used in decision-making is available as quickly as possible. We will see some examples below.

- Code efficiency: we can use ChatGPT to get alternatives to the code we have developed, for example, by specifying that we are looking for a way in which the code is executed as quickly as possible. Each library or function that we use in R is more or less efficient depending on how it uses data and the processes it carries out, and we have seen that multiple options are valid to achieve the same objective in many the cases.
- Functions: if we want to repeat the same process multiple times, the functions are very useful. In the event that we had a code written and needed to transform it into a function, ChatGPT could respond to that need.
- Automation: for questions that have arisen during the course regarding how to read multiple files at once or produce the same graphs for different players/tasks, we can ask ChatGPT to indicate us how to achieve greater automation.

As Sport Scientists, there is a need to have a range of skills and tools, data analysis and RStudio being just two of them. However, as we have seen, in data analysis there are several branches or parts of the process and it is extremely difficult to master each of them due to their rapid expansion, constant novelties and need to keep up with the rest of the technological advancements. ChatGPT can also be used as material to increase our mastery and exploration in advanced areas, as well as to take a step further in parts of the process, as in the following examples.

- Visualization and communication of results: libraries such as ggplot and plotly have great customization capabilities - with the help of ChatGPT we can achieve more advanced visualizations.



- Statistical models: questions to understand the models, or examples that we can run in our RStudio session to experiment with the results. It can also describe results and help in understanding them.

With the subscription version of ChatGPT it is possible to get a key/password to link the account to a specific RStudio package that allows you to use the tool within the R session. In the same way that in the previous module a key/token was created to authorize Rstudio to access Google Sheets, we would use a key/password provided by ChatGPT to use it. If this is an application that may be interesting for certain contexts, we certainly have it at our fingertips.

It is important to know its limitations - it does not replace the work of the Sport Scientist or the Data Analyst, since the elaboration of the question and the development of the work plan are fundamental parts of their job. In addition, as professionals in contact with players, we know the context and the particularities that may be affecting the data - these may be important and may condition the analysis. The decisions that have to be made will be made by the Sport Scientist, so ChatGPT cannot replace this task. In the same way, when we select the method with which we communicate the results, we know our audience, we know the information and the format it will be received in the best way among the staff, players or management, according to their level of understanding, detail of the information, etc. Therefore, these decisions will also have to be made by the Sport Scientist.

Finally, we should keep in mind that it is a model-based tool, and as we saw in previous modules, all models have their corresponding intrinsic error margin. It is important to know the question we are seeking to answer and to be critical of the answers we get, that is, to make sure that the result we are getting is the correct and desired one.



References

Almarie, B.; Teixeira, P. E. P.; Pacheco-Barrios, K.; Rossetti, C. A.; Fregni, F. (2023). The Use of Large Language Models in Science: Opportunities and Challenges. *Principles and Practice of Clinical Research*, 9(1), 1–4. <https://doi.org/10.21801/ppcri.2023.91>

OpenAI. (2021). ChatGPT. <https://www.openai.com/chatgpt/>

