# BASIC PROMPT ENGINEERING WITH CHATGPT: AN INTRODUCTION

## MODULE 4: Evaluation and metrics

## OVERVIEW

In this module, we will concentrate on two critical aspects of prompt engineering: evaluation and metrics. A thorough understanding of how to assess the performance of prompts is vital to ensure their effectiveness and reliability in various applications. In this module, you will develop the tools and knowledge necessary to evaluate and measure the success of your prompt development and crafting efforts.

We will begin by exploring response evaluation criteria. Here, we will develop a simple framework for evaluating ChatGPT responses and changing prompts to increase the quality of responses – the RACCCA framework. We will then explore how using numeric scores allows metrics to be developed that enable the usefulness of prompts to be assessed more effectively.

We will finally explore how prompt response evaluation and metrics can be useful in comparative evaluation of different prompts.

By the end of Module 4, you will be well-versed in evaluation methodologies and metrics, allowing you to assess the performance of your prompt development efforts. This knowledge will enable you to iteratively improve your prompts, ensuring their success and reliability across a wide range of applications.

The module should take between 4 – 7 hours to complete.

## LEARNING OBJECTIVES

By the end of this module you will be able to:

4.1. Describe the usefulness of prompt response evaluation techniques
4.2. Describe the RACCCA approach to evaluating prompt responses
4.3. Provide a metrics-based evaluation of prompt responses
4.4. Iteratively improve prompts to provide better responses against a set of criteria
4.5. Comparatively evaluate different prompts focused on similar outputs

# FLOW

## OVERVIEW

Intro blurb (above)

Intro video:

- This module focuses on the importance of being able to evaluate the effectiveness of a prompt as you both iteratively develop prompts and as you develop your prompt engineering skills
- We will be using the RACCCA framework that was developed for this course. This isn't the only framework, but it's a useful one to know.
- RACCCA stands for Relevance, Accuracy, Completeness, Clarity, Coherence, and Appropriateness.
- At the end of this module you will have a good grasp of prompt evaluation approaches and techniques.
- Just as a heads up, the exercises in this module use techniques that were developed earlier in the course. If you are struggling, you may want to go back and refresh yourself.
- They are also complex, and expect a higher degree of familiarity with ChatGPT than previous modules. It's worth paying especial attention to the instructions.

## ADDITIONAL INFORMATION:

In this module, we will delve into the critical aspects of prompt response evaluation and metrics that will provide you with important knowledge and skills to optimize your prompts for a wide range of applications. The skills learned in this module will be useful for refining your understanding of prompt engineering and enhancing your ability to create effective and reliable prompts.

First, you will explore the significance of prompt response evaluation techniques. You will learn why evaluating ChatGPT-generated responses is important for refining prompt performance and ensuring their effectiveness in various contexts.

Next, you will be introduced to the RACCCA framework for evaluating prompt responses. This is a framework that has been developed for this course. It isn't the only one, but it is a useful starting point. This comprehensive approach, which stands for Relevance, Accuracy, Completeness, Clarity, Coherence, and Appropriateness, will provide you with a structured methodology to systematically assess the quality of generated responses and guide you in modifying prompts accordingly.

As we progress, you will develop an understanding of metrics-based evaluation. You will learn how to use numeric scores to quantitatively measure the success of your prompts. This objective approach will help you to make informed decisions for further improvement and help you effectively assess the usefulness of your prompts—all while remembering that although you are using numbers, your evaluations are still subjective.

This module will also guide you in applying the insights gained from evaluation and metrics to refine your prompts in an iterative manner. This process will ensure that your prompts continually improve and become more effective and reliable over time.

Finally, we will focus on the comparative evaluation of prompts. You will learn how to leverage prompt response evaluation and metrics to compare different prompts focused on similar outputs. This skill will help you make informed decisions when choosing the most effective prompt for a given situation.

By exploring these key aspects, you will gain a deeper understanding of prompt evaluation methodologies and metrics, which will enable you to create and refine prompts that are effective and reliable across various applications.

## EXERCISE: Exploring prompt evaluation and metrics with ChatGPT (15 points)

This exercise is designed to allow you to explore approaches to prompt evaluation and metrics by playing and experimenting with ChatGPT.

Note: In machine learning and machine learning-based prompt-engineering there are a number of formal methods to evaluate the quality of AI responses to questions. These include methods such as BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and METEOR (METEOR (Metric for Evaluation of Translation with Explicit ORdering). In the exercise below, ChatGPT may try and teach you about these – **please note that they are beyond the scope of this class, and you do not need to learn about them here!**

**Exercise:**

- Open a session with ChatGPT, making sure that you are in GPT-4 mode.
- Craft a prompt that asks ChatGPT to teach you about prompt evaluation and metrics when using ChatGPT. At this stage you should have sufficient skill to do this without further instruction.
- If your initial prompt doesn't lead to a useful session with ChatGPT, try different versions of the prompt until you get one that initiates a learning session that is both relevant and at the right level (you will find that some prompts lead to ChatGPT covering material that is not relevant to this course).
- Use your best prompt to learn more about the basics of prompt evaluation and metrics using ChatGPT.
- Paste the prompt you ended up using below.

You are not expected to spend more than 30 minutes on this exercise.

*You will be given full points for successfully completing this exercise. Points may be removed at a future date if it appears that you did not spend as much time and effort as expected on the exercise.*

## EXERCISE: Response Evaluation Criteria (30 points)

There are a number of factors that come into play when evaluating the quality of responses from ChatGPT to a particular prompt. In the context of this course we are focused on developing the professional skills that allow the quality of responses to be assessed so that prompts can be refined and improved.

In this exercise you will explore six specific dimensions of response quality (this isn't an exclusive list, but is a useful one):

1. **Relevance:** The extent to which the response directly addresses the issue or question.
2. **Accuracy:** The degree to which the response provides correct, reliable, and fact-based information.
3. **Completeness:** The degree to which the response covers all essential aspects of the topic or question being asked.
4. **Clarity:** How easily the response can be understood by the intended audience.
5. **Coherence:** The extent to which the response is logically structured and well-organized and flows smoothly from one point to another.
6. **Appropriateness:** How well the response aligns with the intended audience and context and is suitable and respectful in tone and content.

We'll refer to this as the RACCCA framework.

This exercise should take no longer than 1 – 2 hours.

**Exercise:**

The exercise is complex, so please pay attention! It also uses an example of a multi-step prompt template.

### 1. Seed Question

- Come up with a very general question for ChatGPT. For instance "what is a dog" or "How long is a piece of string." Use your imagination!
- Open a chat session with ChatGPT, making sure that you are in GPT-4 mode.
- Ask the question and make a copy of the response.

**2. Response evaluation and prompt refinement**

- Open a **new** session with ChatGPT, making sure that you are in GPT-4 mode.
- Use the following multi-step prompt template to understand where ChatGPT's initial response succeeded or failed, and how improving the prompt can help improve the response:
- **Prompt 1:**
    - Hi ChatGPT. In response to the question [original question] I got the following response: [original response].

      Please assess this in terms of: Relevance (The extent to which the response directly addresses the issue or question); Accuracy (The degree to which the response provides correct, reliable, and fact-based information); Completeness (The degree to which the response covers all essential aspects of the topic or question being asked); Clarity (How easily the response can be understood by the intended audience); Coherence (The extent to which the response is logically structured and well-organized and flows smoothly from one point to another); and Appropriateness (How well the response aligns with the intended audience and context and is suitable and respectful in tone and content)

      [original question]:<br>
      [original response]:
- **Prompt 2:**
    - "Provide three examples of how I could improve the original question to improve the quality of the response"
- **Prompt 3:**
- In a **new** session with ChatGPT, ask the following:
    - [Ask one of the refined prompts provided by ChatGPT in the step above]
- **Prompt 4:**
- In a **new** session with ChatGPT, ask the following:
    - Hi ChatGPT. In response to the question [original question] I got the following response: [original response].

      Please assess this in terms of: Relevance (The extent to which the response directly addresses the issue or question); Accuracy (The degree to which the response provides correct, reliable, and fact-based information); Completeness (The degree to which the response covers all essential aspects of the topic or question being asked); Clarity (How easily the response can be understood by the intended audience); Coherence (The extent to which the response is logically structured and well-organized and flows smoothly from one point to another); and Appropriateness (How well the response aligns with the intended audience and context and is suitable and respectful in tone and content)

[original question]:
[original response]:

- **Prompt 5:**
  - "Provide three examples of how I could improve the original question to improve the quality of the response"
- **Prompt 6:**
- In a **new** session with ChatGPT, ask the following:
  - [Ask one of the refined prompts provided by ChatGPT]
- **Prompt 7:**
- In a **new** session with ChatGPT, ask the following:
  - Hi ChatGPT. In response to the question [original question] I got the following response: [original response].

    Please assess this in terms of: Relevance (The extent to which the response directly addresses the issue or question); Accuracy (The degree to which the response provides correct, reliable, and fact-based information); Completeness (The degree to which the response covers all essential aspects of the topic or question being asked); Clarity (How easily the response can be understood by the intended audience); Coherence (The extent to which the response is logically structured and well-organized and flows smoothly from one point to another); and Appropriateness (How well the response aligns with the intended audience and context and is suitable and respectful in tone and content)

    [original question]:
    [original response]:

### 3. Reflection

Briefly reflect (200 – 300 words) on what you have learned about response evaluation criteria. Do not use ChatGPT for your response!

*You will be given full points for successfully completing this exercise. Points may be removed at a future date if it appears that you did not spend as much time and effort as expected on the exercise -- especially on the reflection.*

## EXERCISE: Response Evaluation Metrics (25 points)

In this exercise we explore how the quality of a response from ChatGPT can be quantified. This is useful in evaluating responses to different prompts (which we will cover in the next exercise). It's also useful in learning how to both develop more effective prompts and assess the quality of the responses that you get.

We will stick with the six criteria used in the RACCCA framework:

- Relevance
- Accuracy
- Completeness
- Clarity
- Coherence
- Appropriateness

Each of these will be assigned a scale from 1 (poor) to 5 (good), and you will use these to evaluate the response of ChatGPT to an initial prompt. You will then refine the prompt until you get as high an overall score as possible.

This exercise is obviously easy to game by just giving ChatGPT full marks first time round. You will learn a lot more though if you are tough with your grading!

This exercise should take no longer than 1 – 2 hours.

*You will be given full points for successfully completing this exercise. Points may be removed at a future date if it appears that you did not spend as much time and effort as expected on the exercise.*

**Exercise**

- Open a new session with ChatGPT, making sure that you are in GPT-4 mode.
- Use the following prompt: "Which are best, apples or oranges?"
- Evaluate ChatGPT's response in terms of the following, and note the aggregate response. Make sure you refer back to the definitions of each above:
    - Relevance (1 – 5)
    - Accuracy (1 – 5)
    - Completeness (1 – 5)
    - Clarity (1 – 5)
    - Coherence (1 – 5)
    - Appropriateness (1 – 5)
- Refine the prompt and iterate to produce responses that lead to a higher score (remember, you are doing the scoring, not ChatGPT). You may find it helpful to start a new ChatGPT

session with ach iteration. You are free to interpret the original prompt in ways that help you refine its usefulness.

- Submit your final prompt and ChatGPT's response below,

*You will be given full points for successfully completing this exercise. Points may be removed at a future date if it appears that you did not spend as much time and effort as expected on the exercise.*

## EXERCISE: Comparative Evaluation of Different Prompts (30 points)

This exercise takes the previous exercise a step further by using the evaluation metrics to compare different versions of a similar prompt.

Even though we will use numeric scores, it's important to remember that these are subjective, and that this is simply a tool to help refine prompts.

We will use the example of asking ChatGPT to help come up with ideas for a futuristic workplace for undergraduate students who are working on assignments.

You will be asked to evaluate the responses to three different prompts using the "RACCCA" scale:

- Relevance (1 – 5)
- Accuracy (1 – 5)
- Completeness (1 – 5)
- Clarity (1 – 5)
- Coherence (1 – 5)
- Appropriateness (1 – 5)

When you apply this scale, think about the type of workplace you would want to spend time in.

The four prompts are:

1. "Design a workspace for undergraduate students to work in assignments in"
2. "Imagine a place that is designed to be as comfortable and as inviting as possible for undergraduate students to hang out in as they are working on assignments"
3. "Provide details of how to design a space where undergraduate students can come and work on assignments without distraction, and where they can be comfortable and focus on their work"
4. "Hi ChatGPT. I would like you to come up with a great design for a space where undergraduate students can come and work on assignments. The space should be incredibly welcoming, not distracting, and extremely comfortable. It should also help undergrads focus on their work and be as productive as possible. Thank you so much!"

This exercise should take no longer than 1 – 2 hours.

**Exercise:**

- Open a new session with ChatGPT, making sure that you are in GPT-4 mode.
    - Provide ChatGPT with prompt 1 from above.
    - Assess the quality of the response using the RACCCA scale and make a note of your assessment.
- Open a **new** session with ChatGPT, making sure that you are in GPT-4 mode.
    - Provide ChatGPT with prompt 2 from above.
    - Assess the quality of the response using the RACCCA scale.
- Open a **new** session with ChatGPT, making sure that you are in GPT-4 mode.
    - Provide ChatGPT with prompt 3 from above.
    - Assess the quality of the response using the RACCCA scale.
- Open a new session with ChatGPT, making sure that you are in GPT-4 mode.
    - Provide ChatGPT with prompt 4 from above.
    - Assess the quality of the response using the RACCCA scale.
- Provide your scores for the response to each prompt below, and write a short (300 – 400) word reflection on how the RACCCA scale helped in evaluating the effectiveness of each prompt, and how this might in turn help you develop better prompts. Do **not** use ChatGPT for your reflection.

*Grading on this exercise will be based on the degree to which you apply and explore the RACCCA framework, and how this is reflected in your reflection. A reflection that includes substantial details about how you applied the RACCCA framework to ChatGPT responses, how this helped (or did not help) in comparing responses to the four prompts, and the usefulness of the framework in developing prompts, will get between 28 – 30 points.*