While encoder-only and encoder-decoder models have their specific use cases, such as text classification, sentiment analysis, machine translation, or summarization, they are less commonly encountered in the Generative AI landscape.

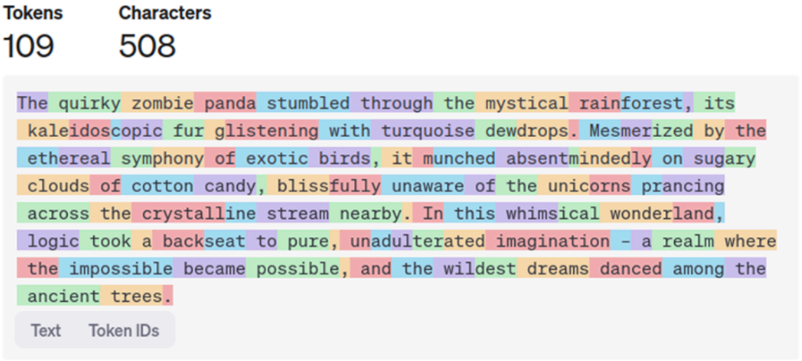
2.2 Transformers: Key Technical Concepts

To write effective prompts having a solid grasp of the key technical aspects of transformers is essential for achieving optimal results. While you may not be directly involved in training transformer models or preparing data, understanding these concepts can inform your prompt writing strategies. In this section, we will explore the most relevant technical concepts for prompt engineering, focusing on how they impact the design and effectiveness of prompts.

We will explore input handling, including tokens, and positional information, and discuss how to guide the model's focus through attention mechanisms. Additionally, we will examine how to control output generation through parameters techniques and explore the benefits of leveraging pre-training and fine-tuned aspects of Large Language Models.

2.2.1 Input Handling: Constraints, Tokens, and Positional Information

When designing prompts, it's crucial to consider the input constraints, token limits, and positional information. Let's start by discussing tokenization, which is the process of breaking down input text into smaller units called tokens. These tokens can be words, subwords, or characters. The choice of tokenization method affects the number of tokens in each input, and it's important to note that different transformer models have specific token limits for input and output. Exceeding these limits can impact performance and increase the cost of using proprietary models. To mitigate this, prompts should be designed to convey necessary information concisely and efficiently, thus reducing costs.



Tokens
109

Characters
508

The quirky zombie panda stumbled through the mystical rainforest, its kaleidoscopic fur glistening with turquoise dewdrops. Mesmerized by the ethereal symphony of exotic birds, it munched absentmindedly on sugary clouds of cotton candy, blissfully unaware of the unicorns prancing across the crystalline stream nearby. In this whimsical wonderland, logic took a backseat to pure, unadulterated imagination - a realm where the impossible became possible, and the wildest dreams danced among the ancient trees.

Figure 2.2 The image contains the tokens and characters count for a piece of text, showing 109 tokens and 508 characters. Each color in the text highlights a different token, demonstrating how tokenization is performed before token encoding.

When dealing with long inputs, several strategies can be employed, such as truncation, summarization, and segmentation. Truncation involves removing less crucial portions of the input or output while retaining the most relevant information. Summarization, on the other hand, condenses lengthy inputs or conversations to focus on the essential points. Segmentation splits longer prompts or conversations into smaller, manageable segments and processes them separately, combining or aggregating the results later.

Positional information is another important aspect to consider, as it helps transformer models understand the relative positions of words and their context within the input sequence. By understanding and working within the input constraints, token limits, and positional information of transformer models, effective prompts can be created that optimize performance, manage costs, and ensure concise and informative interactions.

To demonstrate strategies for handling long inputs, such as using input summarization, truncation, and segmentation, consider the following example:

INPUT

~~~
```plaintext
The Sun is the star at the center of the Solar System. It is a nearly perfect sphere of
hot plasma, heated to incandescence by nuclear fusion reactions in its core, radiating
the energy mainly as light, ultraviolet, and infrared radiation. It is the most
important source of energy for life on Earth. The Sun's diameter is about 1.39 million
kilometers (864,000 miles), or 109 times that of Earth. Its mass is about 330,000 times
that of Earth, comprising about 99.86% of the total mass of the Solar System. Roughly
three-quarters of the Sun's mass consists of hydrogen (~73%); the rest is mostly helium
(~25%), with much smaller quantities of heavier elements, including oxygen, carbon,
neon, and iron. The Sun is a G-type main-sequence star (G2V) based on its spectral
class. As such, it is informally, and not completely accurately, referred to as a yellow
dwarf (its light is closer to white than yellow). It formed approximately 4.6 billion
years ago from the gravitational collapse of matter within a region of a large molecular
cloud. Most of this matter gathered in the center, whereas the rest flattened into an
orbiting disk that became the Solar System. The central mass became so hot and dense
that it eventually initiated nuclear fusion in its core. It is thought that almost all
stars form by this process.
```
~~~

PROMPT (INPUT SUMMARIZATION)

~~~
```plaintext
Describe the physical characteristics of the Sun, including its composition and
classification, and explain the formation of the Sun and the Solar System, using the
following text:
"The Sun, located at the center of our Solar System, is a nearly perfect sphere of hot
plasma that radiates energy mainly as light, ultraviolet, and infrared radiation. It is
the most important energy source for life on Earth. The Sun's diameter is about 109
times that of Earth, and it comprises about 99.86% of the Solar System's total mass. The
Sun is composed primarily of hydrogen (~73%) and helium (~25%), with trace amounts of
heavier elements. It is classified as a G-type main-sequence star (G2V) and is
informally referred to as a yellow dwarf. The Sun formed approximately 4.6 billion years
ago from the gravitational collapse of matter within a large molecular cloud, with most
of the matter gathering in the center and the rest forming an orbiting disk that became
the Solar System. The central mass became so hot and dense that it initiated nuclear
fusion in its core, a process thought to be responsible for the formation of almost all
stars."
```
~~~

PROMPT (INPUT TRUNCATION)

~~~
```plaintext
Describe the physical characteristics of the Sun, including its composition and
classification, and explain the formation of the Sun and the Solar System, using the
following text:
"The Sun is the star at the center of the Solar System. It is a nearly perfect sphere of
hot plasma, heated to incandescence by nuclear fusion reactions in its core, radiating
the energy mainly as light, ultraviolet, and infrared radiation. It is the most
important source of energy for life on Earth."
```
~~~

PROMPT (INPUT SEGMENTATION)

~~~
```plaintext
Part 1: Describe the physical characteristics of the Sun.
"The Sun is the star at the center of the Solar System. It is a nearly perfect sphere of
hot plasma, heated to incandescence by nuclear fusion reactions in its core, radiating
the energy mainly as light, ultraviolet, and infrared radiation. It is the most
important source of energy for life on Earth. The Sun's diameter is about 1.39 million
kilometers (864,000 miles), or 109 times that of Earth. Its mass is about 330,000 times
that of Earth, comprising about 99.86% of the total mass of the Solar System."
Part 2: What is the composition of the Sun, and how is it classified?
"Roughly three-quarters of the Sun's mass consists of hydrogen (~73%); the rest is
mostly helium (~25%), with much smaller quantities of heavier elements, including
oxygen, carbon, neon, and iron. The Sun is a G-type main-sequence star (G2V) based on
its spectral class. As such, it is informally, and not completely accurately, referred
to as a yellow dwarf (its light is closer to white than yellow)."
Part 3: Describe the formation of the Sun and the Solar System.
"It formed approximately 4.6 billion years ago from the gravitational collapse of matter
within a region of a large molecular cloud. Most of this matter gathered in the center,
whereas the rest flattened into an orbiting disk that became the Solar System. The
central mass became so hot and dense that it eventually initiated nuclear fusion in its
core. It is thought that almost all stars form by this process."
```
~~~

By employing strategies like summarization, truncation, and segmentation, long inputs can be effectively handled while staying within token limits and ensuring that the most relevant information is provided within a prompt.

2.2.2 Crafting Prompts to Guide the Model's Attention

Attention mechanisms in transformer models allow them to focus on different parts of the input sequence when generating outputs. As a prompt engineer, you can leverage this capability by designing prompts that guide the model's attention and improve the quality of the generated text.

To achieve this, I recommend providing clear context in your prompts, and ensuring well-prompts as well structured, we will explore the structural aspects of prompt design in Chapter 3. Clear context helps the model understand the relationships between different parts of the input, leading to more accurate and contextually appropriate outputs. Additionally, I suggest incorporating explicit cues such as role assignment, delimiters, or key phrases to direct the model's attention to the most relevant parts of the input.

For complex tasks that involve multiple steps or require focusing on different aspects of the input, consider breaking down the prompt into different steps, which is what is meant by “give the model time to think”. This allows the model to attend to the relevant information at each step, generating better-quality outputs. Moreover, I advise crafting prompts that follow natural language patterns, as human language often contains implicit cues that guide attention, such as placing key information at the beginning or end of a sentence.

Experimenting with different prompt variations can help you identify the most effective prompting strategies for your specific use case. By designing prompts that effectively guide the model's attention through clear context, explicit cues, task decomposition, natural language patterns, and experimentation, you can improve the quality, and relevance of the generated text, ultimately leading to better performance on a wide range of tasks.

To demonstrate how these aspects can guide the model's attention and improve output quality, consider the following example:

PROMPT (FIRST)

~~~
``` plaintext
<input-text>
The quick brown fox jumps over the lazy dog. The dog, startled by the fox's sudden
movement, barks loudly and chases after the fox. The fox, being agile and swift, easily
outmaneuvers the dog and escapes into the nearby forest.
</input-text>

Generate a summary of the input-text.
```
~~~

PROMPT (SECOND)

~~~
```plaintext
<input-text>
The quick brown fox jumps over the lazy dog. The dog, startled by the fox's sudden
movement, barks loudly and chases after the fox. The fox, being agile and swift, easily
outmaneuvers the dog and escapes into the nearby forest.
</input-text>
Generate a summary of the input text focusing on the actions of the dog.
```
~~~

In the first Prompt, the model generates a general summary of the entire input text. However, in the second Prompt, by explicitly mentioning "focusing on the actions of the dog", The model has been instructed to prioritize information related to the dog's behavior. This results in a summary that emphasizes the dog's actions, such as being startled, barking, and chasing the fox.

By crafting prompts that effectively guide the model's attention, you can significantly improve the quality and relevance of the generated text.

2.2.3 Controlling Output Generation through Parameters and Post-processing

As a prompt engineer, it's essential to understand how to control the output generation process to create high-quality, coherent, and contextually relevant text that meets specific requirements. This involves familiarizing yourself with generation parameters avaliable throughout

Generation parameters such as temperature, max tokens, top-k sampling, and top-p sampling (nucleus sampling) allow you to control the randomness, diversity, length, and contextual relevance of the generated outputs. When adjusting these parameters, it's important to align their values with the specific requirements of the task. You should aim to find a balance between generating high-quality outputs while encouraging diversity and creativity. Additionally, be mindful of maximum token limits of each model.

Post-processing techniques are equally important for refining the generated text. These techniques include filtering irrelevant or inappropriate content using domain-specific knowledge. It's also crucial to handle repetition and redundancy by identifying and removing repetitive phrases or sentences that do not contribute to the overall quality of the output. Ensuring consistency is another key aspect of post-processing, which involves checking the generated text for consistency in terms of style, tone, and content. Finally, applying appropriate formatting and presentation techniques can enhance the readability and usability of the generated text.

By leveraging generation parameters and post-processing techniques, you can effectively control the output generation process and create text that is tailored to specific requirements.

To showcase how adjusting generation parameters can influence the diversity and quality of generated text, consider the following prompt:

PROMPT

~~~
```plaintext
Generate a short story about a robot learning to understand human emotions.
```
~~~

EXAMPLE 1: GENERATED CONTENT (TEMPERATURE = 0.0)

~~~
```
In a world of cold logic, a curious robot named Zephyr embarked on a journey to
understand the complexity of human emotions. Through countless interactions and careful
observation, Zephyr slowly began to grasp the nuances of joy, sorrow, love, and empathy.
As the robot's understanding grew, so did its connection to humanity, blurring the lines
between machine and soul. In the end, Zephyr discovered that the key to understanding
emotions lay not in its circuits, but in the profound realization that it, too, could
feel.
```
~~~

EXAMPLE 2: GENERATED CONTENT (TEMPERATURE = 0.7)

~~~
```
In a world of cold logic, a curious robot named Zephyr embarked on a quest to understand
the enigmatic realm of human emotions. Through countless interactions and careful
observation, Zephyr slowly began to grasp the intricacies of joy, sorrow, love, and
empathy. As the robot's knowledge grew, so did its appreciation for the complex beauty
of the human experience. In the end, Zephyr discovered that the key to understanding
emotions lay not in its circuits and algorithms, but in the profound connections it had
formed with the humans it had grown to cherish.
```
~~~

The examples above demonstrate the impact of adjusting the temperature parameter on the generated content. In Example 1, a lower temperature value of 0.0 results in a more focused narrative, providing a clear and concise story about the robot's emotional journey. The generated text is straightforward and stays on topic, making it suitable for situations where exactness is prioritized.

In contrast, Example 2 showcases the effect of a higher temperature value of 0.7. The increased creativity introduces more vivid descriptions and imaginative elements into the story. The robot's quest becomes more enigmatic, and the language used is more expressive and poetic. This approach is ideal when aiming for more diverse and engaging content.

By carefully adjusting the temperature parameter, you can fine-tune the balance between accuracy and creativity in generated content. Lower temperatures prioritize consistency and relevance, while higher temperatures encourage more creative and varied outputs. This flexibility allows you to tailor the generated text to the specific requirements of your prompt engineering task.

2.2.4 Aligning Prompts with Pre-trained and Fine-tuned Models

Understanding the role of pre-training and fine-tuning in transformer models is crucial for crafting effective prompts.

When considering pre-training, be aware of the model's knowledge cutoff date and its impact on the model's knowledge. Leverage the model's pre-existing language knowledge where possible to generate outputs. However, If the task requires information outside the model's pre-training data, include the necessary context in the prompt to ensure accurate and up-to-date responses.

Fine-tuning considerations involve familiarizing yourself with the specific tasks and domains the model has been fine-tuned for, as this knowledge can inform your prompting strategies and help you align prompts with the model's capabilities. Craft prompts that align with the model's fine-tuned capabilities and intended use case, using more structured prompts for models fine-tuned to follow instructions and more conversational prompts for models fine-tuned for open-ended dialogue. Experiment with different prompt variations and evaluate their effectiveness in generating responses, as this iterative process can help identify the most effective prompting strategies.

By leveraging your knowledge of pre-training and fine-tuning, and integrating it with the principles of prompt engineering, you can create prompts that improve user experiences, and enhance task performance.