What is Llama?

WORKING WITH LLAMA 3

Imtihan Ahmed
Machine Learning Engineer

datacamp

Meet the instructor

- Imtihan Ahmed
- Machine Learning Engineer
- Six years experience
- AI at scale

What is Llama 3?

- Summarization

What is Llama 3?

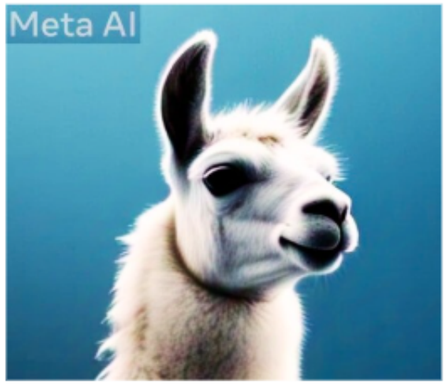
- Summarization

- Data analysis

What is Llama 3?

- Summarization
- Data analysis
- Coding assistant

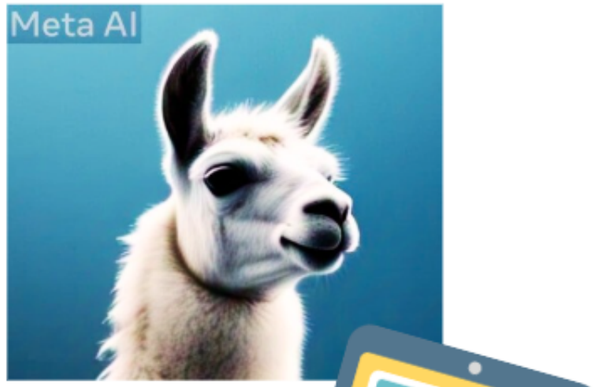
What is Llama 3?



Llama 3:

- Developed by Meta

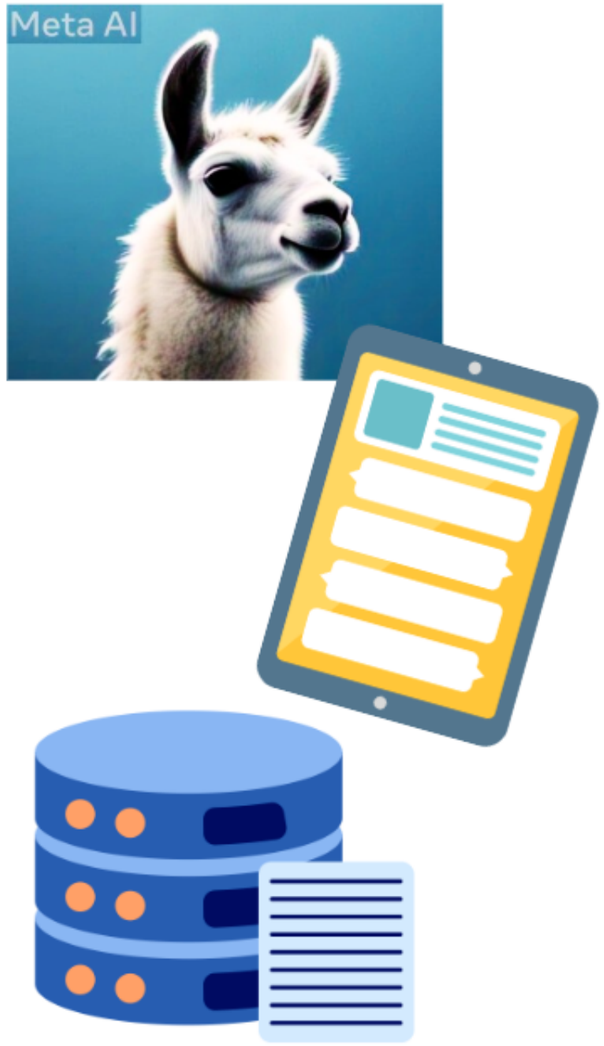
What is Llama 3?



Llama 3:

- Developed by Meta
- Text generation

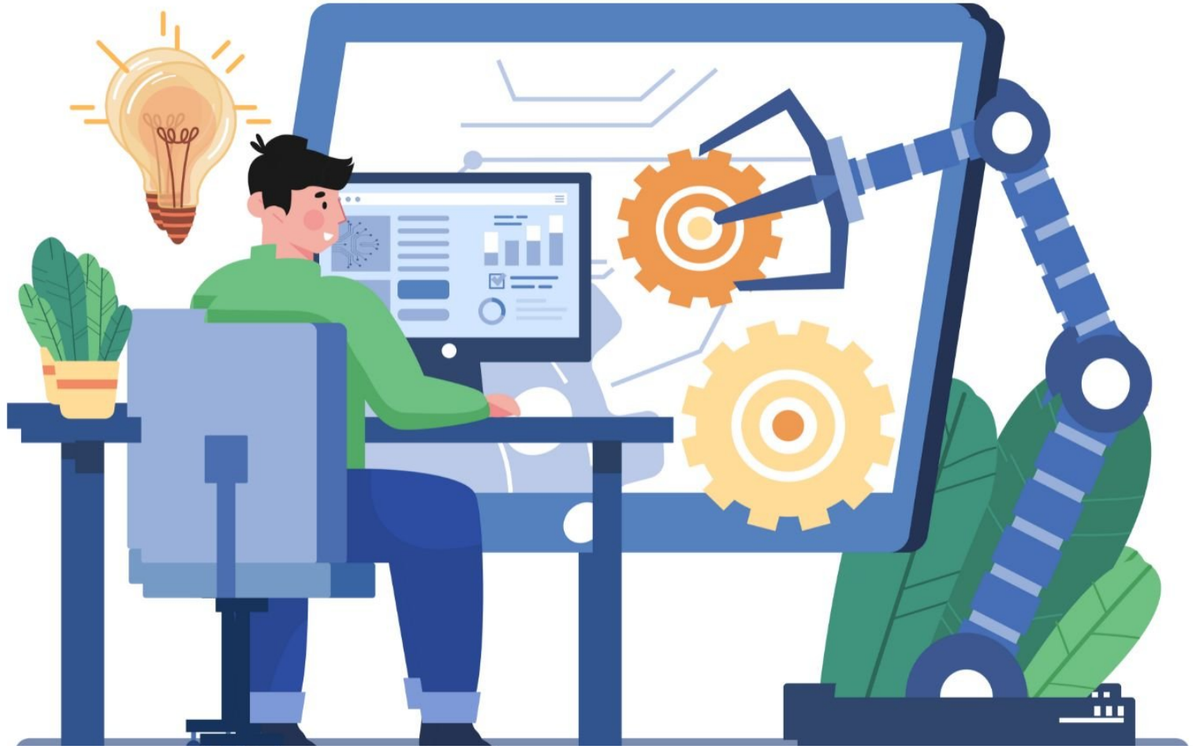
What is Llama 3?



- Developed by Meta
- Text generation
- Trained on massive dataset
  - Equivalent to 2000 Wikipedias
- Open-source

Why run Llama 3 locally

- Example: industrial automation
  - Llama predicting maintenance needs



[1] https://ai.meta.com/blog/aitomatic-built-with-llama/

Why run Llama 3 locally

- Privacy and security

Why run Llama 3 locally

- Privacy and security

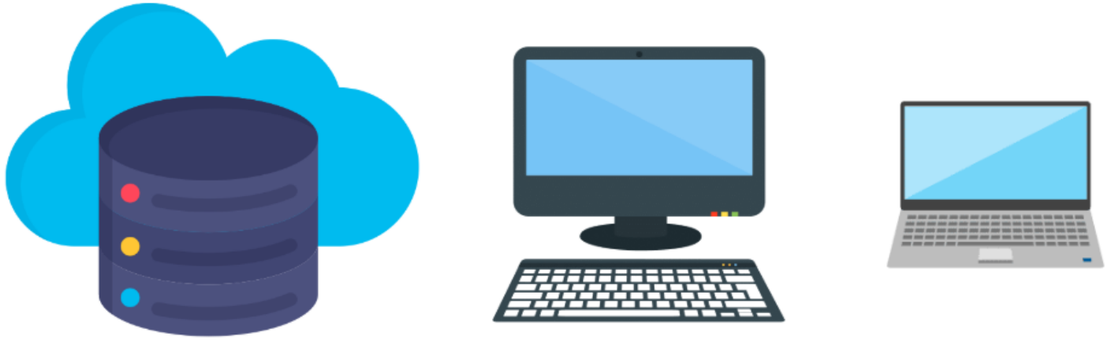
- Cost efficiency

Why run Llama 3 locally

- Privacy and security
- Cost efficiency
- Customization

Using Llama locally

- Available locally when Python is installed



- Can be used through the `llama-cpp-python` library
- Run pip install

```
pip install llama-cpp-python
```

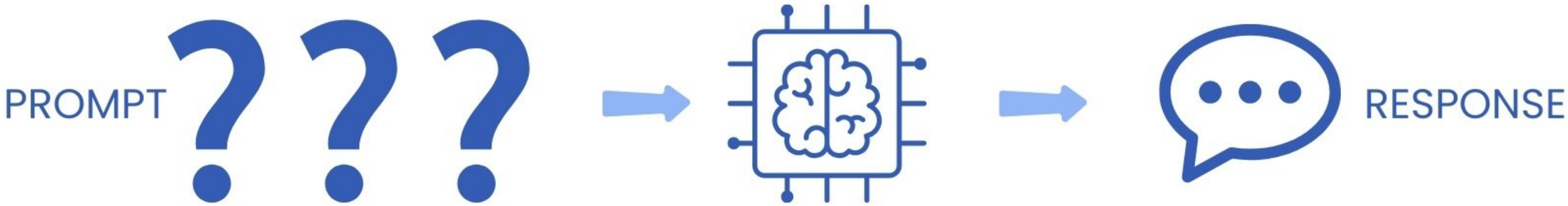
Asking questions to Llama

```
from llama_cpp import Llama
llm = Llama(model_path = "path/to/model.gguf")
output = llm("What are some ways to improve customer retention?")
```

- Import Llama class
- Initialize the LLM
  - Can send prompts and get responses
  - Model in GGUF format
- Pass a question to the LLM

Asking questions to Llama

- Model receives the prompt
- Processes the prompt
- Returns a response



Unpacking the output

- The response is in a dictionary format

```
output
```

```
{'id': 'cmpl-af88304f-97b0-49f5-ba20-db87f86c4068',
 'object': 'text_completion',
 'created': 1715222298,
 'model': './Llama3-gguf-unsloth.Q4_K_M.gguf',
 'choices': [{'text': 'Improving customer retention is crucial for any business, as
 it leads to increased revenue, positive word-of-mouth, and cost savings...'},
 ...]
}
```

Unpacking the output

- Access `"text"` from the `0` element of `"choices"`

```
output["choices"][0]["text"]
```

```
'Improving customer retention is crucial for any business, as
 it leads to increased revenue, positive word-of-mouth, and cost savings.
 Here are some effective ways to boost customer retention:
 1. Personalize the Customer Experience
 Tailor your interactions with customers based on their preferences, behaviors, and
 purchase history. Use data and analytics to create personalized offers,
 recommendations, and communications.
 2. Foster Strong Relationships
 Build strong, human connections with your customers. Train your...'
```

Let's practice!

WORKING WITH LLAMA 3

Tuning Llama 3 parameters

WORKING WITH LLAMA 3

Imtihan Ahmed
Machine Learning Engineer

datacamp

What are parameters for?

```
from llama_cpp import Llama
llm = Llama(model_path="path/to/model.gguf")
output = llm("What are some ways to improve customer retention?")
```

What are parameters for?

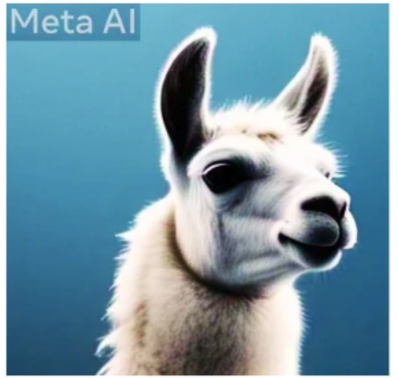
- Example: generating product descriptions

→ Should be factual and concise

→ Should be engaging and creative

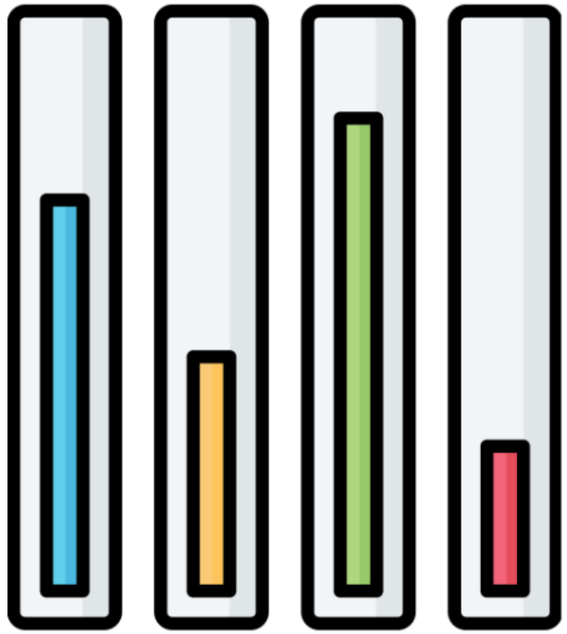
Llama 3 decoding parameters

- Adjust Llama's behavior
- Use **decoding parameters** to match different tones
- Transform raw output into readable text



Llama 3 decoding parameters

- **Temperature:** controls randomness
- **Top-K:** limits token selection to the most probable choices
- **Top-P:** adjusts token selection based on cumulative probability
- **Max tokens:** limits response length



Temperature

- Values usually between 0 and 1
- **Low temperature** (e.g., close to 0):
  - More predictable response

```
A smartwatch with a heart rate monitor, GPS, and a long-lasting batter
for all-day tracking.
```

- **High temperature** (e.g., close to 1):
  - More creative response

```
Your personal fitness coach on your wrist - track every heartbeat, eve
and every adventure without limits.
```

Top-k

- Limits how many of the most likely words Llama can choose from
- **Low k value** (e.g., 1):
  - More predictable response

```
Track fitness, stream music, and receive notifications with our sleek
```

- **High k value** (e.g., 50):
  - More diverse response

```
Experience the future with our cutting-edge smartwatch, featuring fitn
```

Top-p

- Controls the choice of output words based on confidence
- **High top-p value** (e.g., close to 1):
  - More varied responses

```
Stay connected with our sleek smartwatch, featuring fitness tracking,
music, and customizable notifications, perfect for fitness
enthusiasts and busy professionals.
```

- **Low top-p value** (e.g., close to 0):
  - Less variation

```
Smartwatch with fitness tracking and music control, perfect for workouts.
```

Max tokens

- Used to limit response length
- The count of tokens - units of words - in the response
- **Low max_tokens value:**

```
Stay connected with our sleek smartwatch, featuring fitness tracking
and music control.
```

- **High max_tokens value:**

```
Stay connected with our sleek smartwatch, featuring fitness tracking,
music control, customizable notifications, and seamless smartphone
integration. Monitor your health, track your progress, and receive
alerts on your wrist. Perfect for fitness enthusiasts.
```

Combining different parameters

```
llm = Llama(model_path="path/to/model.gguf")


output_concise = llm(
    "Describe an electric car.",
    temperature=0.2,
    top_k=1,
    top_p=0.4,
    max_tokens=20
)
```

```
A fast, eco-friendly electric car with a long range and cutting-edge technology.
```

Combining different parameters

```
output_creative = llm(
    "Describe an electric car.",
    temperature=0.8,
    top_k=10,
    top_p=0.9,
    max_tokens=100
)
```

```
Glide into the future with an electric car that blends speed, luxury,
and sustainability. Silent yet powerful, it redefines the road ...
```

Let's practice!

WORKING WITH LLAMA 3

Assigning chat roles

WORKING WITH LLAMA 3

Imtihan Ahmed
Machine Learning Engineer

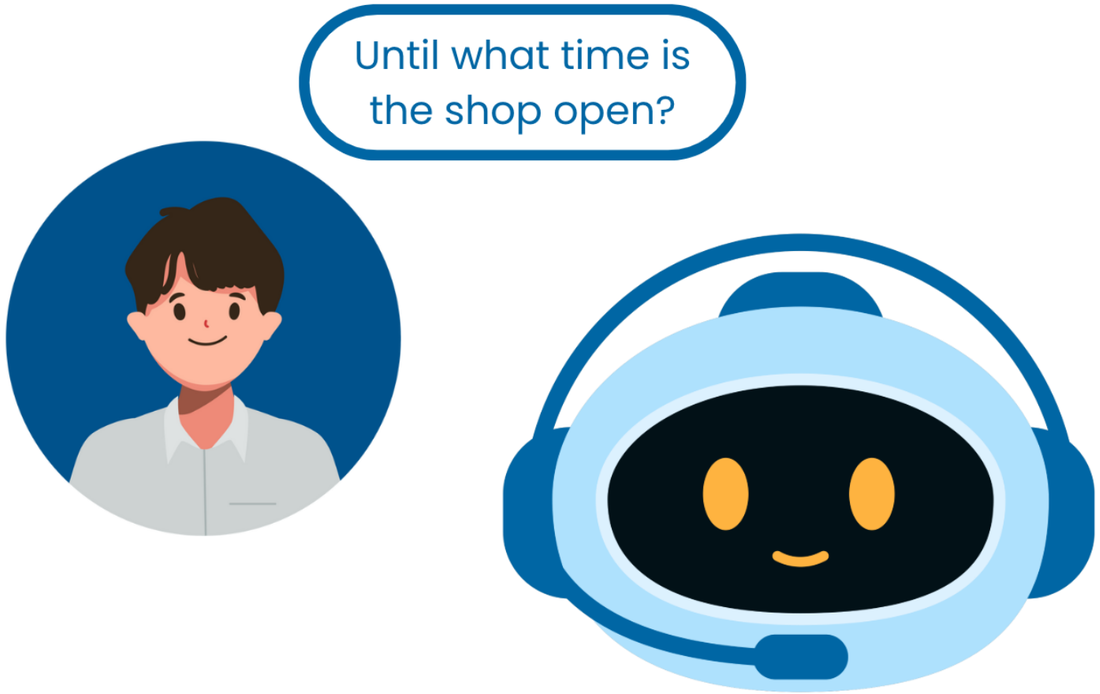
datacamp

Defining roles

- Refining model's tone
- Example: customer support chatbot

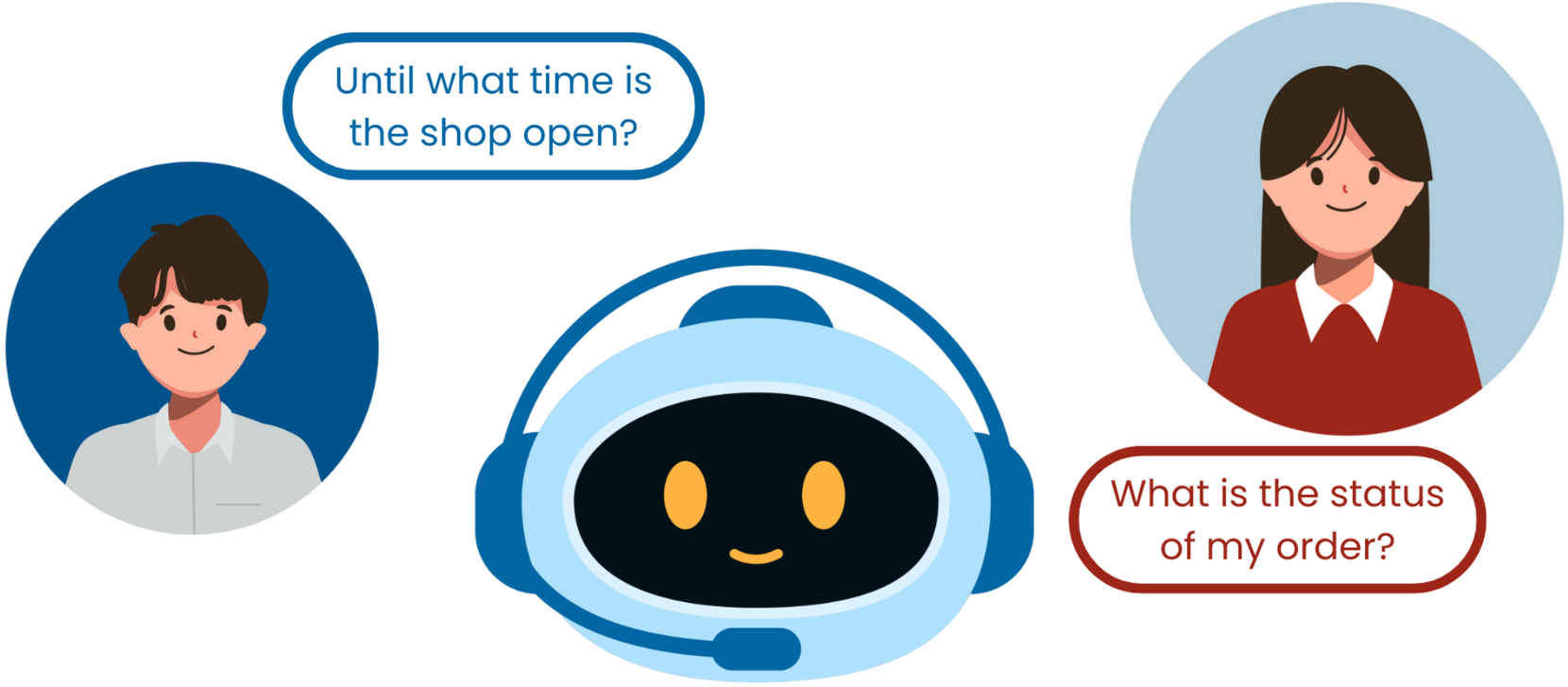
Defining roles

- Refining model's tone
- Example: customer support chatbot



Defining roles

- Refining model's tone
- Example: customer support chatbot



Using roles in chat completion

- Chat roles to guide Llama's responses

- Sets the personality and style

- Represents the person asking the question

Using roles in chat completion

- Sending a structured conversation
- `create_chat_completion()` function

```
from llama_cpp import Llama

llm = Llama(model_path="path/to/model.gguf")

message_list = [...] # This list includes roles

response = llm.create_chat_completion(
        messages = message_list
)
```

The system role

- **System** message: instructions about how model should behave

```
system_message = "You are a business consultant who gives data-driven answers."



message_list = [{
                "role": "system",
                "content": system_message
                }]
```

The user role

- **User** message: the prompt being asked to the model

```
system_message = "You are a business consultant who gives data-driven answers."
user_message = "What are the key factors in a successful marketing strategy?"

message_list = [{"role": "system", "content": system_message},
                {
                 "role": "user",
                 "content": user_message
                 }
                ]
```

Generating the response

```
from llama_cpp import Llama
llm = Llama(model_path="path/to/model.gguf")

system_message = "You are a business consultant who gives data-driven answers."
user_message = "What are the key factors in a successful marketing strategy?"

message_list = [{"role": "system", "content": system_message},
                {"role": "user", "content": user_message}]

response = llm.create_chat_completion(messages = message_list)
print(response)
```

```
{'id': ..., 'object': ..., 'created': ..., 'model': ..., 'choices': [...], ...}
```

The assistant role

```
response["choices"][0]
```

```
 {'index': 0, 'message': {'role': 'assistant', 'content': 'A successful
marketing strategy relies on clear objectives, established through specific,
measurable goals. Understanding the target audience ...'},
'logprobs': None, 'finish_reason': 'length'}
```

The assistant role

```
response["choices"][0]
```

```
 {'index': 0, 'message': {'role': 'assistant', 'content': 'A successful
marketing strategy relies on clear objectives, established through specific,
measurable goals. Understanding the target audience ...'},
'logprobs': None, 'finish_reason': 'length'}
```

The assistant role

```
response["choices"][0]
```

```
 {'index': 0, 'message': {'role': 'assistant', 'content': 'A successful
marketing strategy relies on clear objectives, established through specific,
measurable goals. Understanding the target audience ...'},
'logprobs': None, 'finish_reason': 'length'}
```

The assistant role

```
response["choices"][0]
```

```
 {'index': 0, 'message': {'role': 'assistant', 'content': 'A successful
marketing strategy relies on clear objectives, established through specific,
measurable goals. Understanding the target audience ...'},
'logprobs': None, 'finish_reason': 'length'}
```

```
result['choices'][0]['message']['content']
```

```
'A successful marketing strategy relies on clear objectives, established
through specific, measurable goals. Understanding the target audience ...'
```

Let's practice!

WORKING WITH LLAMA 3