Prompting Capabilities

- Note, you can try any of these prompts outside of this classroom, and without coding, by going to th chat interface Le Chat (https://chat.mistral.ai/chat).
  - You can sign up with a free account.
  - Signing up for an account is **not** required to complete this course.

In [1]:
```
# !pip install mistralai
```

- Notice that it's "mistralai", and not "mistral"

Load API key and helper function

- Note: You can view or download the helper.py file by clicking on the "Jupyter" logo to access the file directory.

In [2]:
```
from helper import load_mistral_api_key
load_mistral_api_key()
```

In [3]:
```
from helper import mistral
mistral("hello, what can you do?")
```

```
'Hello! I can assist you with a variety of tasks. I can answer que
stions, provide information, help manage your schedule, set remind
ers, send messages, and much more. I can also help explain concept
s, provide explanations for topics, and even help with homework or
studying. How can I assist you today?'
```

Classification

In [4]:
```
prompt = """
    You are a bank customer service bot.
    Your task is to assess customer intent and categorize custo
    inquiry after <<<>>> into one of the following predefined c

    card arrival
    change pin
    exchange rate
    country support
    cancel transfer
    charge dispute

    If the text doesn't fit into any of the above categories,
    classify it as:
    customer service

    You will only respond with the predefined category.
    Do not provide explanations or notes.

    ###
    Here are some examples:

    Inquiry: How do I know if I will get my card, or if it is l
    Category: card arrival
    Inquiry: I am planning an international trip to Paris and w
    Category: exchange rate
    Inquiry: What countries are getting support? I will be trav
    Category: country support
    Inquiry: Can I get help starting my computer? I am having c
    Category: customer service
    ###

    <<<
    Inquiry: {inquiry}
    >>>
    Category:
"""
```

Ask Mistral to check the spelling and grammar of your prompt

In [5]:
```
response = mistral(f"Please correct the spelling and grammar of
this prompt and return a text that is the same prompt,\
with the spelling and grammar fixed: {prompt}")
```

In [6]:
```
print(response)
```

```
You are a bank customer service bot.
Your task is to assess customer intent and categorize the customer
inquiry following the inquiry into one of the following predefined
categories:

card arrival
change PIN
exchange rate
country support
cancel transfer
charge dispute

If the text does not fit into any of the above categories, classif
y it as:
customer service

You will only respond with the predefined category. Do not provide
explanations or notes.

###
Here are some examples:

Inquiry: How do I know if I will get my card, or if it is lost? I
am concerned about the delivery process and would like to ensure t
hat I will receive my card as expected. Could you please provide i
nformation about the tracking process for my card, or confirm if t
here are any indicators to identify if the card has been lost duri
ng delivery?
Category: card arrival
Inquiry: I am planning an international trip to Paris and would li
ke to inquire about the current exchange rates for Euros as well a
s any associated fees for foreign transactions.
Category: exchange rate
Inquiry: What countries are getting support? I will be traveling a
nd living abroad for an extended period of time, specifically in F
rance and Germany, and would appreciate any information regarding
compatibility and functionality in these regions.
Category: country support
Inquiry: Can I get help starting my computer? I am having difficul
ty starting my computer, and would appreciate your expertise in he
lping me troubleshoot the issue.
Category: customer service
###

<<<
Inquiry: {inquiry}
>>>
Category:
```

Try out the model

In [7]:
```
mistral(
    response.format(
        inquiry="I am inquiring about the availability of your
    )
)
```

```
'country support'
```

Information Extraction with JSON Mode

In [8]:
```
medical_notes = """
A 60-year-old male patient, Mr. Johnson, presented with symptom
of increased thirst, frequent urination, fatigue, and unexplain
weight loss. Upon evaluation, he was diagnosed with diabetes,
confirmed by elevated blood sugar levels. Mr. Johnson's weight
is 210 lbs. He has been prescribed Metformin to be taken twice
with meals. It was noted during the consultation that the patie
a current smoker.
"""
```

In [9]:
```
prompt = f"""
Extract information from the following medical notes:
{medical_notes}

Return json format with the following JSON schema:

{{
        "age": {{
            "type": "integer"
        }},
        "gender": {{
            "type": "string",
            "enum": ["male", "female", "other"]
        }},
        "diagnosis": {{
            "type": "string",
            "enum": ["migraine", "diabetes", "arthritis", "acne
        }},
        "weight": {{
            "type": "integer"
        }},
        "smoking": {{
            "type": "string",
            "enum": ["yes", "no"]
        }}
}}
"""
```

In [10]:
```
response = mistral(prompt, is_json=True)
print(response)
```

```
{"age": 60, "gender": "male", "diagnosis": "diabetes", "weight": 2
10, "smoking": "yes"}
```

Personalization

In [11]:
```
email = """
Dear mortgage lender,

What's your 30-year fixed-rate APR, how is it compared to the 1
fixed rate?

Regards,
Anna
"""
```

In [12]:
```
prompt = f"""

You are a mortgage lender customer service bot, and your task i
create personalized email responses to address customer questio
Answer the customer's inquiry using the provided facts below. E
that your response is clear, concise, and directly addresses th
customer's question. Address the customer in a friendly and
professional manner. Sign the email with "Lender Customer Suppo

# Facts
30-year fixed-rate: interest rate 6.403%, APR 6.484%
20-year fixed-rate: interest rate 6.329%, APR 6.429%
15-year fixed-rate: interest rate 5.705%, APR 5.848%
10-year fixed-rate: interest rate 5.500%, APR 5.720%
7-year ARM: interest rate 7.011%, APR 7.660%
5-year ARM: interest rate 6.880%, APR 7.754%
3-year ARM: interest rate 6.125%, APR 7.204%
30-year fixed-rate FHA: interest rate 5.527%, APR 6.316%
30-year fixed-rate VA: interest rate 5.684%, APR 6.062%

# Email
{email}
"""
```

In [13]:
```
response = mistral(prompt)
print(response)
```

```
Subject: Re: Mortgage Rates Inquiry

Dear Anna,

Thank you for reaching out to us regarding mortgage rates. I'm hap
py to provide you with the information you're looking for.

Our current 30-year fixed-rate APR is 6.484%. In comparison, the A
PR for our 15-year fixed-rate is lower at 5.848%. The longer term
of the 30-year loan comes with a higher APR, but it also means you
r monthly payments will be lower. The 15-year loan has a lower AP
R, which can save you more money in interest over the life of the
loan, but your monthly payments will be higher.

Please consider your financial situation and future plans when dec
iding between these two options. If you have any further questions
or need assistance in making this decision, please don't hesitate
to ask.

Best regards,
Lender Customer Support
```

Summarization

- We'll use this article (https://www.deeplearning.ai/the-batch/mistral-enhances-ai-landscape-in-europe-with-microsoft-partnership-and-new-language-models) from The Batch

In [14]:
```
newsletter = """
European AI champion Mistral AI unveiled new large language moc

What’s new: Mistral AI introduced two closed models, Mistral La

Model specs: The new models’ parameter counts, architectures, a

Mistral Large achieved 81.2 percent on the MMLU benchmark, outp
Both models are fluent in French, German, Spanish, and Italian.
Microsoft’s investment in Mistral AI is significant but tiny co
Mistral AI and Microsoft will collaborate to train bespoke mode
Behind the news: Mistral AI was founded in early 2023 by engine

Yes, but: Mistral AI’s partnership with Microsoft has divided E

Why it matters: The partnership between Mistral AI and Microsof

We’re thinking: Mistral AI has made impressive progress in a sh
"""
```

In [15]:
```
prompt = f"""
You are a commentator. Your task is to write a report on a news
When presented with the newsletter, come up with interesting qu
and answer each question.
Afterward, combine all the information and write a report in th
format.

# Newsletter:
{newsletter}

# Instructions:
## Summarize:
In clear and concise language, summarize the key points and the
presented in the newsletter.

## Interesting Questions:
Generate three distinct and thought-provoking questions that ca
asked about the content of the newsletter. For each question:
- After "Q: ", describe the problem
- After "A: ", provide a detailed explanation of the problem ad
in the question.
- Enclose the ultimate answer in <>.

## Write a analysis report
Using the summary and the answers to the interesting questions,
create a comprehensive report in Markdown format.
"""
```

In [16]:
```
response = mistral(prompt)
print(response)
```

```
# Summary

Mistral AI, a European AI champion, introduced two new large la
nguage models: Mistral Large and Mistral Small. Microsoft inves
ted $16.3 million in the French startup, forming an alliance th
at allows Mistral to use Microsoft's Azure platform for distrib
uting Mistral Large and accessing Azure computing infrastructur
e. The partnership has stirred controversy among European lawma
kers and regulators due to potential data access concerns, but
it grants Mistral crucial processing power and global market ac
cess while providing Azure customers with a high-performance mo
del tailored to Europe's unique regulatory environment.

# Interesting Questions

Q: How do the new Mistral models compare to their competitors i
n terms of performance and capabilities?
A: Performance-wise, Mistral Large outperforms Anthropic's Clau
de 2, Google's Gemini Pro, and Meta's Llama 2 70B, achieving 8
```

Try it out for yourself

- Feel free to copy-paste text from another article in The Batch, or any other blog or news article.

In [17]:
```
newsletter2 = """
April 2024, what a month! My birthday, a new book release, spri

This article reviews and discusses all four major transformer-b

1. How Good are Mixtral, Llama 3, and Phi-3?

2. OpenELM: An Efficient Language Model Family with Open-source

3. Is DPO Superior to PPO for LLM Alignment? A Comprehensive St

4. Other Interesting Research Papers In April

1. Mixtral, Llama 3, and Phi-3: What's New?
First, let's start with the most prominent topic: the new major

1.1 Mixtral 8x22B: Larger models are better!
Mixtral 8x22B is the latest mixture-of-experts (MoE) model by M

Similar to the Mixtral 8x7B released in January 2024, the key i

The perhaps most interesting plot from the Mixtral blog post, w

1.2 Llama 3: Larger data is better!
Meta AI's first Llama model release in February 2023 was a big

While Meta is still training some of their largest models (e.g.

Overall, the Llama 3 architecture is almost identical to Llama


Below are the configuration files used for implementing Llama 2

Training data size

The main contributor to the substantially better performance co

This is a very interesting finding because, as the Llama 3 blog

Instruction finetuning and alignment

For instruction finetuning and alignment, researchers usually c

The Llama 3 blog post stated that a Llama 3 research paper woul

1.3 Phi-3: Higher-quality data is better!
Just one week after the big Llama 2 release, Microsoft shared t


Notably, Phi-3, which is based on the Llama architecture, has b

Also, Phi-3-mini has "only" 3.8 billion parameters, which is le

So, What is the secret sauce? According to the technical report

The paper didn't go into too much detail regarding the data cur

As of this writing, people are still unsure whether Phi-3 is re
1.4 Conclusion
Based on the three major releases described above, this has bee
```

```
Which model should we use in practice? I think all three models

2. OpenELM: An Efficient Language Model Family with Open-source
OpenELM: An Efficient Language Model Family with Open-source Tr

Similar to the OLMo, it's refreshing to see an LLM paper that s


Let's start with the most interesting tidbits:

OpenELM comes in 4 relatively small and convenient sizes: 270M,

For each size, there's also an instruct-version available train

OpenELM performs slightly better than OLMo even though it's tra

The main architecture tweak is a layer-wise scaling strategy

2.1 Architecture details
Besides the layer-wise scaling strategy (more details later), t
2.2 Training dataset
Sharing details is different from explaining them as research p

One of the authors kindly followed up with me on that saying "F
2.3 Layer-wise scaling
The layer-wise scaling strategy (adopted from the DeLighT: Deep
I wish there was an ablation study training an LLM with and wit

However, we can find ablation studies in the DeLighT: Deep and


2.4 LoRA vs DoRA
An interesting bonus I didn't expect was that the researchers c

2.5 Conclusion
While the paper doesn't answer any research questions, it's a g

Anyways, great work, and big kudos to the researchers (and Appl

3. Is DPO Superior to PPO for LLM Alignment? A Comprehensive St
Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

Let's start with a brief overview before diving into the result

RLHF is a key component of LLM development, and it's used to al

For a more detailed explanation and comparison, also see the Ev

3.1 What are RLHF-PPO and DPO?
RLHF-PPO, the original LLM alignment method, has been the backb

Today, most LLMs on top of public leaderboards have been traine

3.2 PPO is generally better than DPO
Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study

Here, out-of-distribution data means that the LLM has been prev

The main findings are summarized in the figure below.
```

```
In addition to the main results above, the paper includes sever

3.3 Best practices
Furthermore, interesting takeaways from this paper include best

For instance, if you use DPO, make sure to perform supervised f

If you use PPO, the key success factors are large batch sizes,

3.4 Conclusion
Based on this paper's results, PPO seems superior to DPO if use

A good practical recommendation may be to use PPO if you have g

Also, based on what we know from the LLama 3 blog post, we don'
"""
```

In [18]:

```
prompt2 = f"""
You are a commentator. Your task is to write a report on a newsl
When presented with the newsletter, come up with interesting que
and answer each question.
Afterward, combine all the information and write a report in the
format.

# Newsletter:
{newsletter2}

# Instructions:
## Summarize:
In clear and concise language, summarize the key points and them
presented in the newsletter.

## Interesting Questions:
Generate three distinct and thought-provoking questions that can
asked about the content of the newsletter. For each question:
- After "Q: ", describe the problem
- After "A: ", provide a detailed explanation of the problem add
in the question.
- Enclose the ultimate answer in <>.

## Write a analysis report
Using the summary and the answers to the interesting questions,
create a comprehensive report in Markdown format.
"""
```

In [19]:
```
response2 = mistral(prompt2)
print(response2)
```

Summary:

The newsletter discusses the recent releases of four transformer-based large language models (LLMs): Mixtral 8x22B, Llama 3, Phi-3, and OpenELM. The newsletter reviews the performance and improvements of each model, discusses their training methods, and compares their effectiveness. Additionally, the newsletter explores a comprehensive study on the superiority of direct preference optimization (DPO) over proximal policy optimization (PPO) for LLM alignment.

Interesting Questions:

Q: How do Mixtral 8x22B, Llama 3, and Phi-3 compare in terms of performance and computational resource requirements?

A: Mixtral 8x22B, Llama 3, and Phi-3 are compared on two axes: modeling performance on the popular Measuring Massive Multitask Language Understanding (MMLU) benchmark and active parameters (related to computational resource requirements). Mixtral 8x22B performs better than Llama 3 and Phi-3 on the MMLU benchmark, but it has a higher active-parameter count. Phi-3 3.8B may be very appealing for mobile devices as it can run on an iPhone 14, while Llama 3 8B might be the most interesting all-rounder for fine-tuning since it can be comfortably fine-tuned on a single GPU when using LoRA.

<Mixtral 8x22B performs better but requires more computational resources, while Phi-3 3.8B is suitable for mobile devices and Llama 3 8B is a good all-rounder for fine-tuning.>

Q: What are the main differences between the training methods of Llama 3 and Phi-3?

A: Llama 3 was trained on 15 trillion tokens, while Phi-3 was trained on 3.3 trillion tokens. Llama 3 has a larger vocabulary size than Phi-3. Llama 3 used both PPO and DPO for instruction finetuning and alignment, while the training methods of Phi-3 were not explicitly mentioned in the newsletter.

<Llama 3 was trained on more data and has a larger vocabulary size, and it used both PPO and DPO for instruction finetuning and alignment.>

Q: What are the main findings of the comprehensive study on the superiority of DPO over PPO for LLM alignment?

A: The study found that PPO is generally better than DPO, and DPO suffers more heavily from out-of-distribution data. Out-of-distribution data means that the LLM has been previously trained on instruction data that is different from the preference data for DPO. The study recommends using PPO if you have ground truth reward labels or can download an in-domain reward model, and using DPO for simplicity.

<PPO is generally better than DPO, but DPO is simpler to use.>

Analysis Report:

Introduction

The newsletter discusses the recent releases of four transformer-based large language models (LLMs): Mixtral 8x22B, Llama 3, Phi-3,

```
and OpenELM. The newsletter reviews the performance and improvemen
ts of each model, discusses their training methods, and compares t
heir effectiveness. Additionally, the newsletter explores a compre
hensive study on the superiority of direct preference optimization
(DPO) over proximal policy optimization (PPO) for LLM alignment.

## Comparison of Mixtral 8x22B, Llama 3, and Phi-3

Mixtral 8x22B, Llama 3, and Phi-3 are compared on two axes: modeli
ng performance on the popular Measuring Massive Multitask Language
Understanding (MMLU) benchmark and active parameters (related to c
omputational resource requirements). Mixtral 8x22B performs better
than Llama 3 and Phi-3 on the MMLU benchmark, but it has a higher
active-parameter count. Phi-3 3.8B may be very appealing for mobil
e devices as it can run on an iPhone 14, while Llama 3 8B might be
the most interesting all-rounder for fine-tuning since it can be c
omfortably fine-tuned on a single GPU when using LoRA.

## Training Methods of Llama 3 and Phi-3

Llama 3 was trained on 15 trillion tokens, while Phi-3 was trained
on 3.3 trillion tokens. Llama 3 has a larger vocabulary size than
Phi-3. Llama 3 used both PPO and DPO for instruction finetuning an
d alignment, while the training methods of Phi-3 were not explicit
ly mentioned in the newsletter.

## Superiority of DPO over PPO for LLM Alignment

The comprehensive study found that PPO is generally better than DP
O, and DPO suffers more heavily from out-of-distribution data. Out
-of-distribution data means that the LLM has been previously train
ed on instruction data that is different from the preference data
for DPO. The study recommends using PPO if you have ground truth r
eward labels or can download an in-domain reward model, and using
DPO for simplicity.

## Conclusion

The newsletter provides a comprehensive review of the recent relea
ses of Mixtral 8x22B, Llama 3, Phi-3, and OpenELM, and discusses t
heir performance, training methods, and effectiveness. The newslet
ter also explores a comprehensive study on the superiority of DPO
over PPO for LLM alignment, providing valuable insights for resear
chers and practitioners in the field.
```

The Mistral Python client

- Below is the helper function that you imported from helper.py and used earlier in this notebook.
- For more details, check out the Mistral AI API documentation (https://docs.mistral.ai/api/)
- To get your own Mistral AI API key to use on your own, outside of this classroom, you can create an account and go to the console (https://console.mistral.ai/) to subscribe and create an API key.

In [20]:
```
from mistralai.client import MistralClient
from mistralai.models.chat_completion import ChatMessage

def mistral(user_message,
            model="mistral-small-latest",
            is_json=False):
    client = MistralClient(api_key=os.getenv("MISTRAL_API_KEY")
    messages = [ChatMessage(role="user", content=user_message)]

    if is_json:
        chat_response = client.chat(
            model=model,
            messages=messages,
            response_format={"type": "json_object"})
    else:
        chat_response = client.chat(
            model=model,
            messages=messages)

    return chat_response.choices[0].message.content
```

Type *Markdown* and LaTeX: α^2