OpenAI ChatGPT

What is gpt-3.5-turbo?

- Also known as ChatGPT API
- 10x cheaper per 1k tokens
- Optimized for Chat
- Can not be fine-tuned (As at 1 March 2023)

What else is new

- Previous models used completions or instructions
- This model uses a back and forth interaction between a user and an assistant
- Can still do single-turn tasks
- Less moderated than ChatGPT
- Data is not used by OpenAI

Update your libraries

- Do it now!!

Assumptions

- You will have reviewed the completions part of this course
- Light on code – more of a transition
- Update your code libraries !!

Old and New Code

Old Python Code

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.Completion.create(
                model="text-davinci-003",
                prompt="Say this is a test\n")
```

New Python Code

```
import os
import openai

openai.api_key = os.getenv("OPENAI_API_KEY")

response = openai.ChatCompletion.create(
            model="gpt-3.5-turbo",
            messages=[
                  {"role": "system", "content": "You are a helpful assistant."},
                  {"role": "user", "content": "Who won the world series in 2020?"},
                  {"role": "assistant", "content": "The Los Angeles Dodgers."},
                  {"role": "user", "content": "Where was it played?"}
            ]
)
```

Old Response


```
{
  "object": "text_completion",
  "model": "text-davinci:003",
  "choices": [
    {
      "text": "This is indeed a test"
    }
  ]
}
```


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

New Response

```
{
'object': 'chat.completion',
'model': 'gpt-3.5-turbo',
'choices': [{
        'message': {
            'role': 'assistant',
            'content': 'Arlington Texas at the Globe Life Field.'}
        }]
}
```

```
response["choices"][0]["message"]["content"]
```

vs

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

CURL example

```
curl https://api.openai.com/v1/chat/completions \
  -H 'Content-Type: application/json' \
  -H 'Authorization: Bearer YOUR_API_KEY' \
  -d '{
  "model": "gpt-3.5-turbo",
  "messages": [{"role": "user", "content": "Hello!"}]
}'
```

NOTE : https://api.openai.com/v1/chat/completions

Managing Tokens

Counting tokens

- 4096 tokens for both models
- Previously prompt + completion
- Now the SUM of ALL the content within the message
  - Plus a small overhead for formatting
- Allow for the overhead
- Default max_tokens is inf

Chat Parameters

```
messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers."},
    {"role": "user", "content": "Where was it played?"}
]
```

Roles

- system
- user
- assistant

- system
- user, assistant, user, assistant, user

System Role

```
messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers."},
        {"role": "user", "content": "Where was it played?"}
]
```

System Role

You are ChatGPT, a large language model trained by OpenAI. Answer as concisely as possible. Knowledge cutoff: {knowledge_cutoff} Current date: {current_date}

User / Assistant

```
messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers."},
    {"role": "user", "content": "Where was it played?"}
]
```

Order matters

- Move system role to the end of the list

- System : You are Charles Darwin.
- User : Who are you?
- Assistant : I am Charles Darwin

Order matters

- Move system role to the end of the list

- System : You are Charles Darwin.
- User : Who are you?
- User : Who are you?
- Assistant : I am Charles Darwin

Order matters

- Move system role to the end of the list

- System : You are Charles Darwin.
- User : Who are you?
- User : Who are you?
- User : Who are you?
- Assistant : I am an AI language model created by OpenAI.

Order matters

- Move system role to the end of the list

- User : Who are you?
- User : Who are you?
- User : Who are you?
- System : You are Charles Darwin.
- Assistant : I am Charles Darwin.

Making a chatbot

```
messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers."},
        {"role": "user", "content": "Where was it played?"}
]
```

Making a chatbot

```
messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers."},
        {"role": "user", "content": "Where was it played?"}
]
```

- Globe Life Field in Arlington

Making a chatbot

```
messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who won the world series in 2020?"},
        {"role": "assistant", "content": "The Los Angeles Dodgers."},
        {"role": "user", "content": "Where was it played?"}
        {"role": "assistant", "content": "Globe Life Field in Arlington."},
        {"role": "user", "content": "When was it played?"}
]
```

- October 20 to October 27

Introducing Knowledge

```
messages=[
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Who won the world series in 2020?"},
    {"role": "assistant", "content": "The Los Angeles Dodgers."},
    {"role": "user", "content": "The 2020 World Series was the championship series of Major League Baseball's 2020 season. The 116th World Series was a best-of-seven-playoff between the American League (AL) champion Tampa Bay Rays and the National League (NL) champion Los Angeles Dodgers. The Dodgers defeated the Rays to win the series in six games for their first championship since 1988.\n\nWhere was it played?"}
]
```

Temperature

- Range 0 to 2
- Anything above 0.8 will make the output more random
- Lower values like 0.2 will make it more focused and deterministic
- A lot of people are using 0.5 or 0.6

Streaming

- Not documented – but a valid setting
- stream = true
- Easiest with a library that supports streaming

Streaming

- async request
- Read the response stream

- Each line starts with “data:” (which you need to remove)
- The rest of the message will be
  - “[DONE]”
  - JSON string

Streaming


```
{"id":"",
    "object":"chat.completion.chunk",
    "model":"gpt-3.5-turbo-0301",
    "choices":[
       {
          "delta":{ "content": " hello" },
          "index":0,
          "finish_reason":null
       }]
}
```


- response["choices"][0]["delta"]["content"]

Other settings

- top_p
- n
- stop
- presence_penalty
- frequency_penalty
- logit_bias
- user

Missing Settings

- prompt
- suffix
- logprobs
- echo
- best_of

Getting the Best Results

- https://community.openai.com
- OpenAI discord server

- Blog on https://blinkdata.com

- https://github.com/openai/openai-cookbook/blob/main/techniques_to_improve_reliability.md