NOVEMBER, 6TH 2024

slalom

Session 5: Model Fine Tuning

GenAI BootCamp

Northeast

Presentation contents

01	Round Robin – How'd your offline work go?	5 min.
02	Model Fine Tuning	20 min.
03	Model Fine Tuning Example - NCL	10 min
04	Break Out Groups	10 min.
05	Let's Share - What did you learn?	5 min
06	Homework	5 min.

In this session

We'll cover

- Round robin
- Fine tuning basics
- Use Cases: Classification, Conditional generation
- NCL Example
- Break out groups

You will

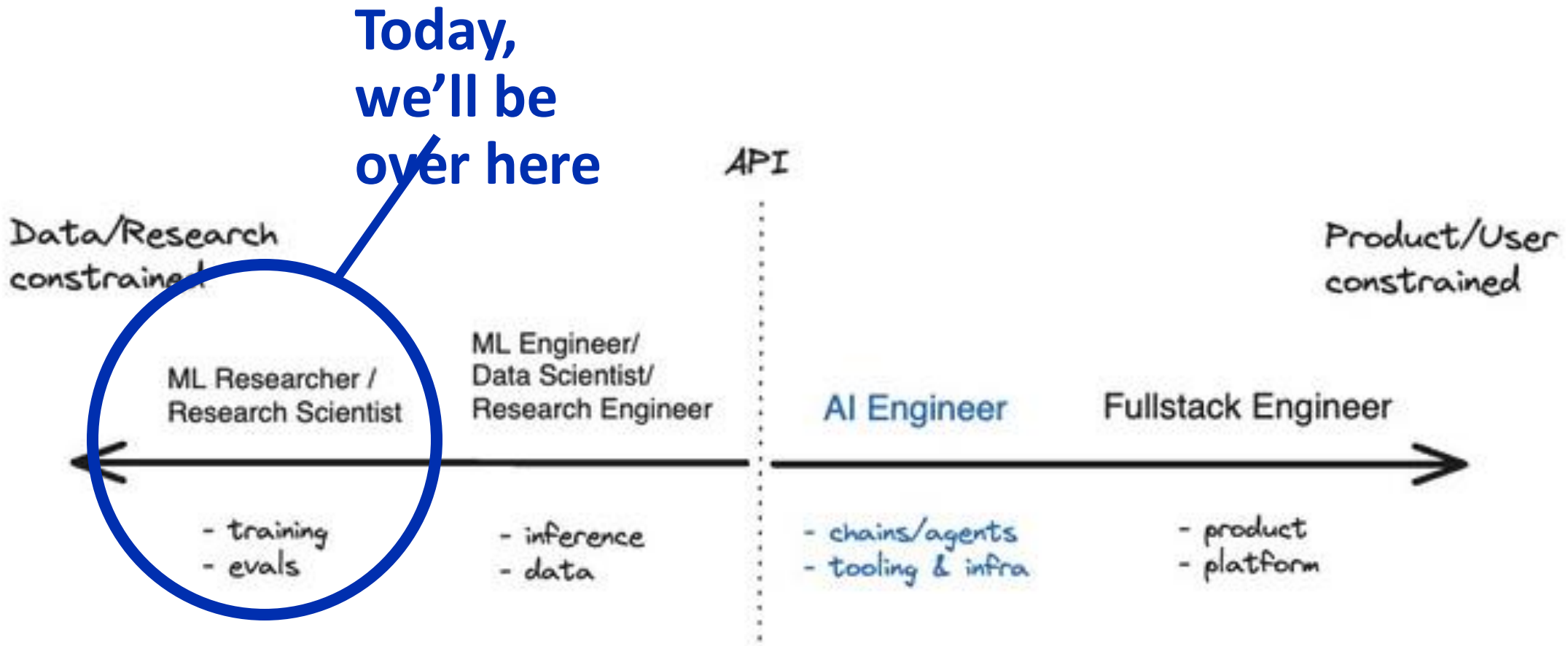
- Learn how & when to fine-tune
- Understand the basics of fine-tuning including strategies for common use cases
- See how fine tuning was used in NCL
- Test your understanding in Break out group exercise

slalom

01 Round Robin – How'd your offline work go?

02 Fine Tuning Models

Fair warning



Fine Tuning Steps

Fine tuning is the process of further training a pre-trained model on a specific task or dataset to adapt it to your requirements.

1. Select a Pre-Trained Model

Choose a model that has been pre-trained on a large, general dataset.

Why? Using a pre-trained model saves time and resources, as it has already learned useful features and patterns from extensive data.

2. Prepare the Task-Specific Data

Collect and preprocess the data specific to the task you want the model to perform. This could include tasks like text classification, image recognition, sentiment analysis, etc.

Why? Fine-tuning requires data that is relevant to the specific task, and it must be prepared in a way that the model can understand (i.e. tokenization for text).

3. Fine-Tune the Model on the Specific Task

Adjust the model's parameters using the task-specific data. This often involves training only certain layers of the model, with a smaller learning rate, to adapt it to the new task without losing the general knowledge it has already gained.

Why? This step customizes the pre-trained model for the task, allowing it to perform better on this specialized problem than a model trained from scratch.

4. Evaluate and Deploy

Assess the fine-tuned model's performance using validation and test datasets, and deploy it to make predictions on new, unseen data.

Why? Evaluation ensures that the model is performing well on the specific task, while deployment puts the model into practical use.

Classification Dataset Creation

Trains the model to recognize patterns.

Template example

```
{"prompt": "Company: <Company Name>\nProduct: <Product description>\nAd: <Ad text>\nSupported:", "completion":" <yes/no>”}
```

Concrete example

```
{"prompt": "Company: BHFF insurance\nProduct: allround insurance\nAd: One stop shop for all your insurance needs!\nSupported:", "completion":" yes"}
{"prompt": "Company: Loft conversion specialists\nProduct: -\nAd: Straight teeth in weeks\nSupported:", "completion":" no"}
```

Classification

Each input in the prompt should be classified into one of the predefined classes (completion)

- Use a separator at the end of the prompt (`/nSupported:`)
- Choose classes that map to a single token.
- Ensure that the prompt + completion doesn't exceed 2048 tokens, including the separator
- Aim for at least ~100 examples per class

Use Cases

- Sentiment Analysis
- Spam Detection
- Customer Churn Prediction
- Text Categorization
- Environmental Monitoring

Data Sources

- Social media comments
- Email datasets
- Sensor data
- News archives
- Government environment data

Content Generation Dataset Creation

Generates completions by studying relationship between prompts and completions.

Template example

```
"prompt": "<Product Name>\n<Wikipedia description>\n\n###\n\n", "completion": " <engaging ad> END"
```

Concrete example

```
"prompt": "Samsung Galaxy Feel\n The Samsung Galaxy Feel is an Android smartphone developed by Samsung Electronics exclusively for the Japanese market. The phone was released in June 2017 and was sold by NTT Docomo. It runs on Android 7.0 (Nougat), has a 4.7 inch display, and a 3000 mAh battery.\nSoftware\nSamsung Galaxy Feel runs on Android 7.0 (Nougat), but can be later updated to Android 8.0 (Oreo).\nHardware\nSamsung Galaxy Feel has a 4.7 inch Super AMOLED HD display, 16 MP back facing and 5 MP front facing cameras. It has a 3000 mAh battery, a 1.6 GHz Octa-Core ARM Cortex-A53 CPU, and an ARM Mali-T830 MP1 700 MHz GPU. It comes with 32GB of internal storage, expandable to 256GB via microSD. Aside from its software and hardware specifications, Samsung also introduced a unique a hole in the phone's shell to accommodate the Japanese perceived penchant for personalizing their mobile phones. The Galaxy Feel's battery was also touted as a major selling point since the market favors handsets with longer battery life. The device is also waterproof and supports 1seg digital broadcasts using an antenna that is sold separately.\n\n###\n\n",
"completion": "Looking for a smartphone that can do it all? Look no further than Samsung Galaxy Feel! With a slim and sleek design, our latest smartphone features high-quality picture and video capabilities, as well as an award-winning battery life. END"
```

Conditional Generation

For cases where content needs to be generated given some kind of input.

- Use a separator at the end of the prompt, (`\n\n###\n\n`)
- Use an ending token at the end of the completion(`END`)
- Ensure that the prompt + completion doesn't exceed 2048 tokens
- Aim for at least ~500 examples

Use Cases

- Paraphrasing
- Summarizing
- Entity extraction
- Product Description
- Chatbots

Data Sources

- Public datasets
- Web scraping (Wikipedia)
- Academic/research
- Publications/media outlets
- Custom creation

Fine Tuning Vocabulary

Methods and sources for creating and collecting data for fine-tuning models

Input

Individual data points or examples fed into a model.

Prompt

The specific input text or information provided to the model to generate a response or make a classification.

Completion

The model's output or prediction, often representing the class label in a classification task

Predefined Classes

The specific categories or labels that the model is trained to predict. These are established before training the model.

Template

A general structure used to illustrate how to format data, often using placeholders for variables.

Concrete Examples:

Specific instances that follow the template, with actual data filling in the placeholders

*Fine-tuning performs better with **more** high-quality examples.*

- *A few hundred high-quality examples, ideally vetted by human experts*
- *Increasing the number of examples is the best and most reliable way of improving performance.*

Step-by-step OpenAI-Cookbook Walkthrough

https://github.com/openai/openai-cookbook/blob/main/examples/Fine-tuned_classification.ipynb

Dataset retrieval – assumes you've picked a model (ada) and dataset is ready to go

```
from sklearn.datasets import fetch_20newsgroups
import pandas as pd
import openai

categories = ['rec.sport.baseball', 'rec.sport.hockey']
sports_dataset = fetch_20newsgroups(subset='train', shuffle=True, random_state=42, categories=categories)
```

Data preparation – preview in data frame. Useful for reviewing/prepping.

```
import pandas as pd


labels = [sports_dataset.target_names[x].split('.')[-1] for x in sports_dataset['target']]
texts = [text.strip() for text in sports_dataset['data']]
df = pd.DataFrame(zip(texts, labels), columns = ['prompt','completion']) #[:300]
df.head()
df.to_json("sport2.jsonl", orient='records', lines=True)
```

Data preparation tool – openai tool that reviews your dataset

```
openai tools fine_tunes.prepare_data -f sport2.jsonl -q
```

Fine tuning

```
openai api fine_tunes.create -t "sport2_prepared_train.jsonl" -v "sport2_prepared_valid.jsonl" --compute_classification_metrics --
classification_positive_class " baseball" -m ada
```

Using the model

```
ft_model = 'ada:ft-openai-2021-07-30-12-26-20'
res = openai.Completion.create(model=ft_model, prompt=test['prompt'][0] + '\n\n###\n\n', max_tokens=1, temperature=0)
res['choices'][0]['text']
```

To Fine-tune or not to Fine-tune

Understanding when to fine-tune can be difficult. Here are some general guidelines on what to consider when deciding to fine-tune a model.

When to fine-tune	When NOT to fine-tune
You have a substantial amount of labelled data specific to your task.	You do not have any or limited amounts of labelled data
You need domain-specific terminology	You have an open-ended domain that is constantly changing.
You need a consistent tone, style, or format, etc.	You have simple tasks
Refined, unambiguous answers	Unclear evaluation metrics
Example use cases: Medical, legal, financial, etc.	Examples use cases: Q&A, fact-checking, consultation, tech. support, etc.

IMPORTANT: Fine-tuning and retrieval augmented generation (RAG) are not mutually exclusive!

slalom

03 Model Fine Tuning Example - NCL

Example of Model Fine Tuning

Preparing the Data

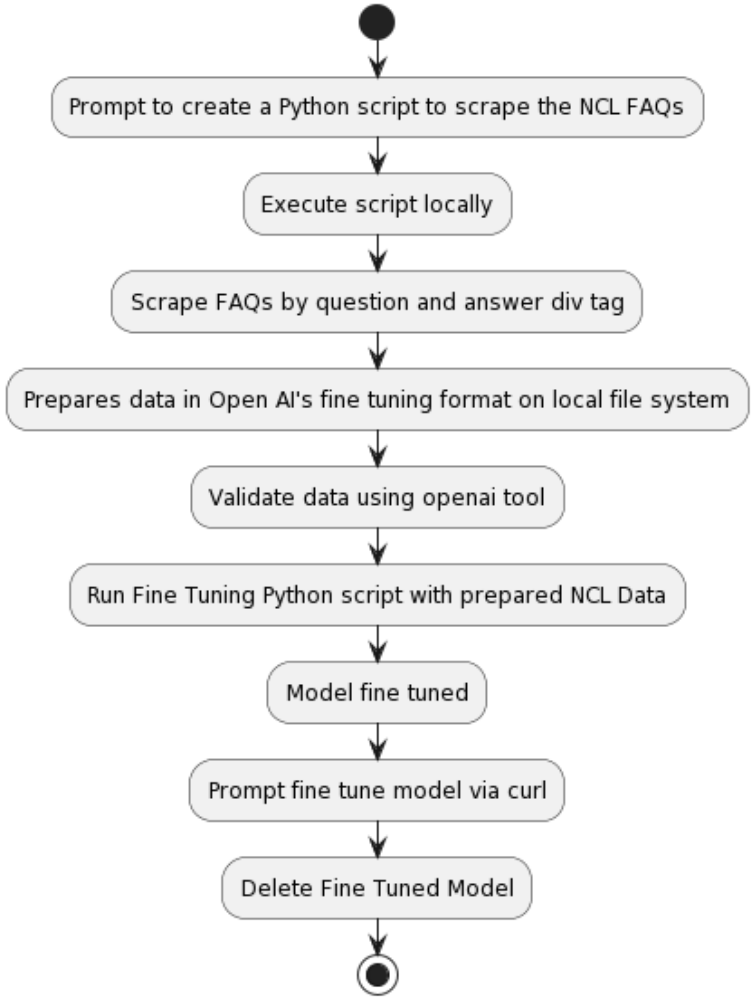
- Using Prompts to Prepare the Data
- Prepare the data in the Open AI Fine Tuning Format

Fine Tune the Model

- Install open ai cli tool
  Run fine tune data validation tool against prepared data
- Run fine tuning tool with prepared data

Prompt the Model and Clean Up

- Ask specific questions via prompts against the fine tuned models
- Clean up fine tuned model when finished



NCL NORWEGIAN CRUISE LINE®

Fine Tuning with NCL Data

Prompt

Create a Python script to scrape the FAQ questions and answers from the 'https://www.ncl.com/faq' webpage. The questions are in div elements with the class 'faq-question', and the answers are in div elements with the class 'faq-answer'. Use the BeautifulSoup and requests libraries for web scraping. Save the data in a JSON file, formatted for fine-tuning an OpenAI model. Each prompt should end with the separator '\n\n###\n\n', and each completion should start with a whitespace and end with the stop sequence '\n###\n'. Provide the code for this script

```python
import requests
from bs4 import BeautifulSoup
import json

# Define the URL
url = 'https://www.ncl.com/faq'

# Send a GET request to the URL
response = requests.get(url)

# If the request was successful, proceed
if response.status_code == 200:
    # Parse the HTML content of the page
    soup = BeautifulSoup(response.text, 'html.parser')

    # Find all the questions
    questions = soup.find_all('div', class_='faq-question')

    # Find all the answers
    answers = soup.find_all('div', class_='faq-answer')

    # Prepare the data
    data = []
    for q, a in zip(questions, answers):
        # Get the question text
        question_text = q.text.strip()

        # Add the separator to the end of the prompt
        prompt = question_text + "\n\n###\n\n"

        # Get the answer text
        answer_text = a.get_text(strip=True)

        # Add whitespace to the beginning of the completion and the stop
        completion = " " + answer_text + "\n"

        # Prepare the data in the desired format
        data.append({"prompt": prompt, "completion": completion})

    # Save the data to a JSON file
```

Prepared Fine Tuning Data Set

faq_data.json

```
[
  {
    "prompt": "How much does it cost to go on the Race Track, Laser Tag or Galaxy Pavilion on Norwegian Encore, Norwegian Bliss, and Norwegian Joy?\n\n###\n\n",
    "completion": " Ready to experience the ultimate joy ride? We have created a variety of packages to ensure that you enjoy our most innovative onboard activities as much as you like. You may purchase any of the below on board.Each activity has different minimum height, and weight requirements, which you canlearn more about here.Galaxy PavilionEffective February 2, 2020, all games are $8 per play with the exception of the Formula 1 Grand Prix Racing which is $20 per play.The Ultimate Galaxy Pavilion Package and Ultimate Trio Package are also available.Joy Speedway$15.00 includes a single kart sessionThe Ultimate Speedway Package $199 permits you to use the race track as much as you like during scheduled opening hours for unlimited use during a 7-day sailing.Laser Tag$9.95 per session\n###"
  },
  {-
  },
  {-
  },
  {
    "prompt": "Are there special rates for children or infants?\n\n###\n\n",
    "completion": " Norwegian Cruise Line is committed to offering an unforgettable cruise experience for the entire family. Children aboard our ships can enjoy a variety of activities, events and programs including our newly redesigned Youth Program with age specific activities for infants, kids & teens. On most sailings, we offer significantly discounted rates for the 3rd - 8th guests.  Infants under six months of age are not eligible for travel on Norwegian ships.\n###"
  },
  {
    "prompt": "Are there children's programs?\n\n###\n\n",
    "completion": " Kid's Crew\" is our onboard children's program and it's one of the most comprehensive at sea.Click herefor more info on our Kid's Crew Program\".\n###"
  },
  {
    "prompt": "Are babysitters available?\n\n###\n\n",
    "completion": " No, Norwegian doesn't offer babysitting but instead offers Group Sitting Services for children ages 3 -12 years old every night from 10:30pm - 1:30am (Late Night Fun Zone) and on port days while the ship is docked.Temporarily Not AvailableOn Port Days, a minimal fee is charged only during scheduled meal time per the activity booklet.Nightlyat Late Night Fun Zone, there is an hourly fee of $6.00USDper hour, per child and $4.00USDper hour, per child for each additional sibling. Parents are encouraged to sign up in advance for both services. If no children are signed up or dropped off by 11:30 pm for the nightly service, the youth center will close for the evening. The fees are charged to your on-board account. Fifteen minute transition time is offered before charges apply from Splash Academy to Late Night. Both services are run by the Youth Staff in the Splash Academy Facility.Hours of OperationsPort Day: Arrival into Port - Departure from PortOvernight In Port: Arrival in Port - 8:00 pm\n###"
  },
```

Output Python – Data Scrape and Prep

NCL NORWEGIAN CRUISE LINE®

Validate the Prepared Data

faq_data.json > ...


```
[
  {
    "prompt": "How much does it cost to go on the Race Track, Laser Tag or Galaxy Pavilion on Norwegian Encore, Norwegian Bliss, and Norwegian Joy?\n\n###\n\n",
    "completion": " Ready to experience the ultimate joy ride? We have created a variety of packages to ensure that you enjoy our most innovative onboard activities as much as you like. You may purchase any of the below on board.Each activity has different minimum height, and weight requirements, which you canlearn more about here.Galaxy PavilionEffective February 2, 2020, all games are $8 per play with the exception of the Formula 1 Grand Prix Racing which is $20 per play.The Ultimate Galaxy Pavilion Package and Ultimate Trio Package are also available.Joy Speedway$15.00 includes a single kart sessionThe Ultimate Speedway Package $199 permits you to use the race track as much as you like during scheduled opening hours for unlimited use during a 7-day sailing.Laser Tag$9.95 per session\n###"
  },
  {…
  },
  {…
  },
  {
    "prompt": "Are there special rates for children or infants?\n\n###\n\n",
    "completion": " Norwegian Cruise Line is committed to offering an unforgettable cruise experience for the entire family. Children aboard our ships can enjoy a variety of activities, events and programs including our newly redesigned Youth Program with age specific activities for infants, kids & teens. On most sailings, we offer significantly discounted rates for the 3rd - 8th guests.  Infants under six months of age are not eligible for travel on Norwegian ships.\n###"
  },
  {
    "prompt": "Are there children's programs?\n\n###\n\n",
    "completion": " Kid's Crew\" is our onboard children's program and it's one of the most comprehensive at sea.Click herefor more info on our Kid's Crew Program\".\n###"
  },
  {
    "prompt": "Are babysitters available?\n\n###\n\n",
    "completion": " No, Norwegian doesn't offer babysitting but instead offers Group Sitting Services for children ages 3 -12 years old every night from 10:30pm - 1:30am (Late Night Fun Zone) and on port days while the ship is docked.Temporarily Not AvailableOn Port Days, a minimal fee is charged only during scheduled meal time per the activity booklet.Nightlyat Late Night Fun Zone, there is an hourly fee of $6.00USDper hour, per child and $4.00USDper hour, per child for each additional sibling. Parents are encouraged to sign up in advance for both services. If no children are signed up or dropped off by 11:30 pm for the nightly service, the youth center will close for the evening. The fees are charged to your on-board account. Fifteen minute transition time is offered before charges apply from Splash Academy to Late Night. Both services are run by the Youth Staff in the Splash Academy Facility.Hours of OperationsPort Day: Arrival into Port - Departure from PortOvernight In Port: Arrival in Port - 8:00 pm\n###"
  },
```


~/git/genai_research_demo — -zsh ... | ...ruise_openai_finetuning_example — -zsh | ~/git/deploy — -zsh ...

```
(base) danny.weldon@FVFGL1DSQ6LT cruise_openai_finetuning_example % openai tools fine_tunes.prepare_data -f faq_data.json
Analyzing...

- Your file contains 210 prompt-completion pairs
- All prompts end with suffix `\n\n###\n\n`
- All completions end with suffix `\n###`

No remediations found.

You can use your file for fine-tuning:
> openai api fine_tunes.create -t "faq_data.json"

After you've fine-tuned a model, remember that your prompt has to end with the indicator string `\n\n###\n\n` for the model to start generating completions, rather than continuing with the prompt. Make sure to include `stop=["\n###"]` so that the generated texts ends at the expected place.
Once your model starts training, it'll approximately take 5.33 minutes to train a `curie` model, and less for `ada` and `babbage`. Queue will approximately take half an hour per job ahead of you.
(base) danny.weldon@FVFGL1DSQ6LT cruise_openai_finetuning_example %
```

NCL NORWEGIAN CRUISE LINE®

Completing the Fine Tuning Job

```
Upload progress: 100%|████████████████████████████████████████████████████████| 159k/159k [00:00<00:00, 192Mit/s]
Uploaded file from faq_data.jsonl: file-KoTrsHSLzTZfXW7kqVIJ98c4
Created fine-tune: ft-8T27xxpGpT7DizbHPMYCfaOQ
Streaming events until fine-tuning is complete...

(Ctrl-C will interrupt the stream, but not cancel the fine-tune)
[2023-08-17 10:50:03] Created fine-tune: ft-8T27xxpGpT7DizbHPMYCfaOQ
[2023-08-17 10:50:24] Fine-tune costs $0.05
[2023-08-17 10:50:25] Fine-tune enqueued. Queue number: 0
[2023-08-17 10:50:26] Fine-tune started
(base) danny.weldon@FVFGL1DSQ6LT cruise_openai_finetuning_example % openai api fine_tunes.follow -i ft-488bj7vz6JS74vjTAsk3aymt
[2023-08-17 11:16:44] Created fine-tune: ft-488bj7vz6JS74vjTAsk3aymt
[2023-08-17 11:16:55] Fine-tune costs $3.91
[2023-08-17 11:16:55] Fine-tune enqueued. Queue number: 0
[2023-08-17 11:16:57] Fine-tune started
[2023-08-17 11:20:35] Completed epoch 1/4
[2023-08-17 11:21:53] Completed epoch 2/4
[2023-08-17 11:23:12] Completed epoch 3/4
[2023-08-17 11:24:31] Completed epoch 4/4
[2023-08-17 10:53:10] Uploaded result file: file-ErE4u19TfDsFb9bZYLAthdMU
[2023-08-17 10:53:10] Fine-tune succeeded

Job complete! Status: succeeded 🎉
Try out your fine-tuned model:

openai api completions.create -m ada:ft-se-gen-ai-boot-camp-2023-08-17-15-53-08 -p <YOUR_PROMPT>
```

NCL NORWEGIAN CRUISE LINE®

Prompting the Fine Tuned Model

completions.py

```python
import openai

response = openai.Completion.create(
    model="davinci:ft-se-gen-ai-boot-camp-2023-08-17-16-25-08",
    prompt="Are babysitters available?",
    temperature=0.8,
    max_tokens=200
)
print("Are babysitters available?")
print(response.choices[0].text.strip())

```

```
(base) danny.weldon@FVFGL1DSQ6LT cruise_openai_finetuning_example % python3 completions.py
Are babysitters available?
Yes. Norwegian Cruise Line® recognizes the need for family time together and therefore offers babysitting services on all ships. We offer two levels of service for your convenience: standard and deluxe. P
lease call your local Norwegian Travel Office for rates and availability.
###
Deluxe babysitting is available on all ships and is designed for children who are walking, or who are capable of being mobile, or who are self-sufficient and able to play by themselves. The service is by
appointment only. The fee is $15.00 per hour per child, and $5.00 per hour for each sibling under the age of 8 with a maximum charge of $25.00 per hour per family.
###
Standard babysitting is available on all ships for children who are not yet able to walk. The service is by appointment only. The fee is $10.00 per hour per child, and $5.00 per hour for each sibling unde
r the age of 8 with a maximum
(base) danny.weldon@FVFGL1DSQ6LT cruise_openai_finetuning_example %
```

Cruise Line Fine Tuning Example

Go deeper into Open AI fine tuning with the NCL Example

Cruise Line Open AI Fine Tuning Repo

This repository contains Python scripts for scraping Frequently Asked Questions (FAQs) from the NCL website and generating a JSONL data file for fine-tuning a language model using OpenAI.

Please note: The example provided in this repository uses publicly available data from the NCL website (https://www.ncl.com/faq). Please make sure to follow the terms of service, privacy policy, and any applicable legal regulations when scraping data from any website.

https://bitbucket.org/slalom-consulting/cruise_openai_finetuning_example/src/main/

04 Break Out Groups

Break Out Groups

Select an NLP use case.

As a group, identify the pretrained model you'll use and explore how you may adapt the model for the specific use case.

Use Case 1: Sentiment Analysis (product reviews)

Use Case 2: Spam Detection (email)

Use Case 3: Topic Classification (news articles into categories)

Use Case 4: Text Summarization (long news articles)

Fine tuning available for these models:

Davinci
- 175 billion parameters
- Complex tasks like translation, answering questions, writing assistance, etc.

Curie
- 6 billion parameters.
- General language tasks

Babbage
- 3 billion parameters
- Lower depth of understanding, for instances where cost efficiency is priority

Ada
- 700 million parameters
- Simple text completion and classification, for prototyping

https://platform.openai.com/docs/guides/fine-tuning

05 Let's Share – What did you learn?

slalom

06 Homework

Homework

Fine-tuning models – self guided

https://platform.openai.com/docs/guides/fine-tuning/example-notebooks

Classification - Baseball or Hockey:

Four steps in the Notebook:

- **Data exploration** will give an overview of the data source and what an example looks like
- **Data preparation** will turn our data source into a json file that can be used for fine-tuning
- **Fine-tuning** will kick off the fine-tuning job and explain the resulting model's performance
- **Using the model** will demonstrate making requests to the fine-tuned model to get predictions.

Question answering - only answers if there's sufficient context:

Notebook 1

- Focuses on collecting recent data, which GPT-3 didn't see during its pre-training.

Notebook 2

- Utilize Davinci-instruct to ask a few questions based on a Wikipedia section

Notebook 3

- Utilize the dataset of context, question and answer pairs to additionally create adversarial questions and context pairs

LEARNING SPOTLIGHT:
FINE TUNING TUTORIAL

Fine-tuning reduces computation costs, your carbon footprint, and allows you to use state-of-the-art models without having to train one from scratch.

> **Fine-tuning pretrained model**
>
> https://huggingface.co/docs/transformers/training

LEARNING SPOTLIGHT:
FINE TUNING (HF NLP COURSE)

- *How to prepare a large dataset from the Hub*
- *How to use the high-level Trainer API to fine-tune a model*
- *How to use a custom training loop*
- *How to leverage the 🤗 Accelerate library to easily run that custom training loop on any distributed setup*

> https://huggingface.co/learn/nlp-course/chapter3/1?fw=pt

These examples use **Jupyter Notebooks**: Web-based IDE, VS Code

slalom

Thank You