Finetuning data: compare to pretraining and basic preparation ¶

In [ ]:
```
import jsonlines
import itertools
import pandas as pd
from pprint import pprint

import datasets
from datasets import load_dataset
```

Look at pretraining data set

Sorry, "The Pile" dataset is currently relocating to a new home and so we can't show you the same example that is in the video. Here is another dataset, the "Common Crawl" (https://huggingface.co/datasets/c4) dataset.

In [ ]:
```
#pretrained_dataset = load_dataset("EleutherAI/pile", split="train", streaming=True)

pretrained_dataset = load_dataset("c4", "en", split="train", streaming=True)
```

In [ ]:
```
n = 5
print("Pretrained dataset:")
top_n = itertools.islice(pretrained_dataset, n)
for i in top_n:
  print(i)
```

Contrast with company finetuning dataset you will be using

In [ ]:
```
filename = "lamini_docs.jsonl"
instruction_dataset_df = pd.read_json(filename, lines=True)
instruction_dataset_df
```

Various ways of formatting your data

In [ ]:
```
examples = instruction_dataset_df.to_dict()
text = examples["question"][0] + examples["answer"][0]
text
```

In [ ]:
```
if "question" in examples and "answer" in examples:
  text = examples["question"][0] + examples["answer"][0]
elif "instruction" in examples and "response" in examples:
  text = examples["instruction"][0] + examples["response"][0]
elif "input" in examples and "output" in examples:
  text = examples["input"][0] + examples["output"][0]
else:
  text = examples["text"][0]
```

In [ ]:
```
prompt_template_qa = """### Question:
{question}

### Answer:
{answer}"""
```

In [ ]:
```
question = examples["question"][0]
answer = examples["answer"][0]

text_with_prompt_template = prompt_template_qa.format(question=question, answer=answe
text_with_prompt_template
```

In [ ]:
```
prompt_template_q = """### Question:
{question}

### Answer:"""
```

In [ ]:
```
num_examples = len(examples["question"])
finetuning_dataset_text_only = []
finetuning_dataset_question_answer = []
for i in range(num_examples):
  question = examples["question"][i]
  answer = examples["answer"][i]

  text_with_prompt_template_qa = prompt_template_qa.format(question=question, answer=
  finetuning_dataset_text_only.append({"text": text_with_prompt_template_qa})

  text_with_prompt_template_q = prompt_template_q.format(question=question)
  finetuning_dataset_question_answer.append({"question": text_with_prompt_template_q,
```

In [ ]:
```
pprint(finetuning_dataset_text_only[0])
```

In [ ]:
```
pprint(finetuning_dataset_question_answer[0])
```

Common ways of storing your data

In [ ]:
```
with jsonlines.open(f'lamini_docs_processed.jsonl', 'w') as writer:
    writer.write_all(finetuning_dataset_question_answer)
```

In [ ]:
```
finetuning_dataset_name = "lamini/lamini_docs"
finetuning_dataset = load_dataset(finetuning_dataset_name)
print(finetuning_dataset)
```