

Processing the dataset takes about 1 minute on an A100 GPU and 6 minutes on an M3 MacBook Air:

```
100% |██████████| 110/110 [01:05<00:00, 1.68it/s]
```

Let's verify that the responses have been correctly added to the `test_set` dictionary by examining one of the entries:

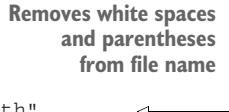
```
print(test_data[0])
```

The output shows that the `model_response` has been added correctly:

```
{'instruction': 'Rewrite the sentence using a simile.',
 'input': 'The car is very fast.',
 'output': 'The car is as fast as lightning.',
 'model_response': 'The car is as fast as a bullet.'}
```

Finally, we save the model as `gpt2-medium355M-sft.pth` file to be able to reuse it in future projects:

```
import re
file_name = f"{re.sub(r'[ ()]', '', CHOOSE_MODEL)}-sft.pth"
torch.save(model.state_dict(), file_name)
print(f"Model saved as {file_name}")
```



The saved model can then be loaded via `model.load_state_dict(torch.load("gpt2-medium355M-sft.pth"))`.

7.8 Evaluating the fine-tuned LLM

Previously, we judged the performance of an instruction-fine-tuned model by looking at its responses on three examples of the test set. While this gives us a rough idea of how well the model performs, this method does not scale well to larger amounts of responses. So, we implement a method to automate the response evaluation of the fine-tuned LLM using another, larger LLM, as highlighted in figure 7.19.

To evaluate test set responses in an automated fashion, we utilize an existing instruction-fine-tuned 8-billion-parameter Llama 3 model developed by Meta AI. This model can be run locally using the open source Ollama application (<https://ollama.com>).

NOTE Ollama is an efficient application for running LLMs on a laptop. It serves as a wrapper around the open source llama.cpp library (<https://github.com/ggerganov/llama.cpp>), which implements LLMs in pure C/C++ to maximize efficiency. However, Ollama is only a tool for generating text using LLMs (inference) and does not support training or fine-tuning LLMs.

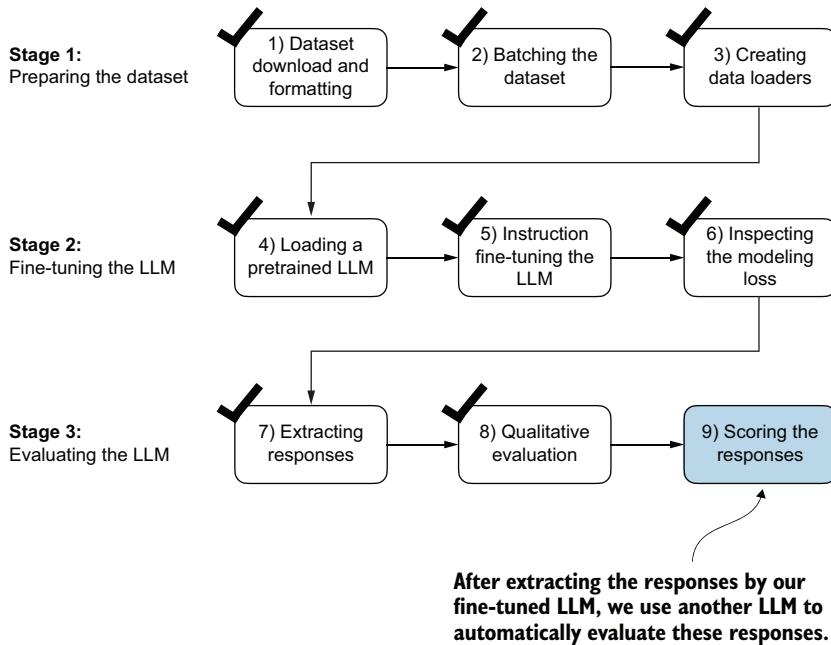


Figure 7.19 The three-stage process for instruction fine-tuning the LLM. In this last step of the instruction-fine-tuning pipeline, we implement a method to quantify the performance of the fine-tuned model by scoring the responses it generated for the test.

Using larger LLMs via web APIs

The 8-billion-parameter Llama 3 model is a very capable LLM that runs locally. However, it's not as capable as large proprietary LLMs such as GPT-4 offered by OpenAI. For readers interested in exploring how to utilize GPT-4 through the OpenAI API to assess generated model responses, an optional code notebook is available within the supplementary materials accompanying this book at <https://mng.bz/BgEv>.

To execute the following code, install Ollama by visiting <https://ollama.com> and follow the provided instructions for your operating system:

- *For macOS and Windows users*—Open the downloaded Ollama application. If prompted to install command-line usage, select Yes.
- *For Linux users*—Use the installation command available on the Ollama website.

Before implementing the model evaluation code, let's first download the Llama 3 model and verify that Ollama is functioning correctly by using it from the command-line terminal. To use Ollama from the command line, you must either start the Ollama application or run `ollama serve` in a separate terminal, as shown in figure 7.20.

First option: make sure to start ollama in a separate terminal via the `ollama serve` command.

```
(base) +> ollama serve
2024/06/06 20:53:14 routes.go:1067: INFO server config env="map[OLLAMA_DEBUG=false OLLAMA_FLASH_ATTENTION=false OLLAMA_HOST= OLLAMA_KEEP_ALIVE= OLLAMA_LLM_LIBRAYR= OLLAMA_MODEL_NAME= OLLAMA_MODEL_PATH= OLLAMA_NUM_PARALLEL=1 OLLAMA_ORIGINS=[http://localhost https://localhost: http://localhost: https://127.0.0.1 https://127.0.0.1: https://127.0.0.1: http://0.0.0.0: http://0.0.0.0: http://sebastian-ollama run llama3 - ollama - ollama run llama3 - 80x24 _OLLAMA_TMRW=0]
time=2024-06-06 last login: Thu Jun 6 20:53:18 on pts/0
obs: 57
time=2024-06-06 >>> What do llamas eat?
used blogs : Llamas are herbivores, which means they primarily eat plants and time=2024-06-06 plant-based foods. Their diet typically consists of:
ng on 127.0.0.1:80
time=2024-06-06 1. Grasses: Llamas love to graze on grasses, including tall grasses, short
    grasses, and even weeds.
    713290/rum/2024-06-06 2. Leaves: They enjoy munching on leaves from trees and shrubs, like oak,
    time=2024-06-06 maple, and willow.
    LLM library 3. Hay: Llamas often eat hay as a staple in their diet, which can include
    time=2024-06-06 alfalfa, timothy, or oat hay.
    computer 4. Grains: Some llamas may receive grains like oats, barley, or corn as
    time=2024-06-06 15-20 G Part of their feed
[GIN] 2024/06-06 5. Fruits and veggies: While not essential to their diet, llamas might
    enjoy treats like apples, carrots, or sweet potatoes.
    6. Minerals: Llamas need access to minerals like salt, calcium, and
    phosphorus to maintain good health.

In the wild, llamas would typically roam free in grasslands, meadows, or
forest edges, where they could forage for their favorite foods. In
captivity, llama owners often provide a mix of these foods to ensure their
animals receive a balanced diet.
```

Then run `ollama run llama3` to download and use the 8-billion-parameter Llama 3 model.

Second option: if you are using macOS, you can also start the ollama application and make sure it is running in the background instead of running `ollama serve`.

Jun 6 8:54PM

● ● ● sebastian — ollama run llama3 — ollama — ollama run llama3 — 80x24

```
Last Logon: Thu Jun 6 20:53:18 on pts/001
(bash) + ollama run llama3
>>> What do Llamas eat?
Llamas are herbivores, which means they primarily eat plants and plant-based foods. Their diet typically consists of:

1. Grasses: Llamas love to graze on grasses, including tall grasses, short grasses, and leafy weeds.
2. Leaves: They enjoy munching on leaves from trees and shrubs, like oak, maple, and willow.
3. Hay: Llamas often eat hay as a staple in their diet, which can include alfalfa, timothy grass, or oat hay.
4. Grains: Some llamas may receive grains like oats, barley, or corn as part of their feed.
5. Fruits and Vegetables: While not essential to their diet, llamas might enjoy treats like apples, carrots, or sweet potatoes.
6. Minerals: Llamas need access to minerals like salt, calcium, and phosphorus to maintain good health.

In the wild, llamas would typically roam free in grasslands, meadows, or forest edges, where they could forage for their favorite foods. In captivity, llama owners often provide a mix of these foods to ensure their animals receive a balanced diet.
```

Figure 7.20 Two options for running Ollama. The left panel illustrates starting Ollama using `ollama serve`. The right panel shows a second option in macOS, running the Ollama application in the background instead of using the `ollama serve` command to start the application.

With the Ollama application or `ollama serve` running in a different terminal, execute the following command on the command line (not in a Python session) to try out the 8-billion-parameter Llama 3 model:

ollama run llama3

The first time you execute this command, this model, which takes up 4.7 GB of storage space, will be automatically downloaded. The output looks like the following:

```
pulling manifest
pulling 6a0746alec1a... 100% |██████████| 4.7 GB
pulling 4fa551d4f938... 100% |██████████| 12 KB
pulling 8ab4849b038c... 100% |██████████| 254 B
pulling 577073ffcc6c... 100% |██████████| 110 B
pulling 3f8eb4ada87fa... 100% |██████████| 485 B
verifying sha256 digest
writing manifest
removing any unused layers
success
```