Works Cited

🤗 *Evaluate*. https://huggingface.co/docs/evaluate/index. Accessed 23 Apr. 2025.

🤗 *Transformers*. https://huggingface.co/docs/evaluate/transformers_integrations. Accessed 23 Apr. 2025.

(23) Enhance Document Management with AI: Extract Insights from PDFs Using Python, Llama 3, and Ollama | LinkedIn. https://www.linkedin.com/pulse/enhance-document-management-ai-extract-insights-from-pdfs-le-sueur-kfd5f/. Accessed 17 Apr. 2025.

(23) LinkedIn. https://www.linkedin.com/pulse/fine-tune-your-ai-ollama-model-files-step-by-step-tutorial-ayres-hfenf/. Accessed 17 Apr. 2025.

A Quick Tour. https://huggingface.co/docs/evaluate/a_quick_tour. Accessed 25 Apr. 2025.

Accelerator. https://huggingface.co/docs/accelerate/v1.6.0/en/package_reference/accelerator#accelerate.Accelerator.backward. Accessed 22 Apr. 2025.

Accuracy - a Hugging Face Space by Evaluate-Metric. https://huggingface.co/spaces/evaluate-metric/accuracy. Accessed 22 Apr. 2025.

All Releases - The Go Programming Language. https://go.dev/dl/. Accessed 17 Apr. 2025.

Apache License, Version 2.0. https://www.apache.org/licenses/LICENSE-2.0. Accessed 4 Apr. 2025.

Axolotl-Ai-Cloud/Axolotl. 2023. Axolotl AI, 18 Apr. 2025. *GitHub*, https://github.com/axolotl-ai-cloud/axolotl.

BERT Score - a Hugging Face Space by Evaluate-Metric. https://huggingface.co/spaces/evaluate-metric/bertscore. Accessed 22 Apr. 2025.

Bilinear — PyTorch 2.7 Documentation. https://pytorch.org/docs/stable/generated/torch.nn.Bilinear.html. Accessed 25 Apr. 2025.

Build Your Own LLM with LLM Fine-Tuning on macOS Using MLX | by (Λx.x)Eranga | Effectz.AI | Medium. https://medium.com/rahasak/fine-tuning-llms-on-macos-using-mlx-and-run-with-ollama-182a20f1fd2c. Accessed 25 Apr. 2025.

“Built-in Functions.” *Python Documentation*, https://docs.python.org/3/library/functions.html. Accessed 18 Apr. 2025.

Causal Language Modeling. https://huggingface.co/docs/transformers/tasks/language_modeling. Accessed 23 Apr. 2025.

Choosing a Metric for Your Task. https://huggingface.co/docs/evaluate/choosing_a_metric. Accessed 23 Apr. 2025.

Chroma | 🦜🔗 LangChain. https://python.langchain.com/docs/integrations/retrievers/self_query/chroma_self_query/. Accessed 17 Apr. 2025.

Creating and Sharing a New Evaluation. https://huggingface.co/docs/evaluate/creating_and_sharing. Accessed 23 Apr. 2025.

“Csv — CSV File Reading and Writing.” *Python Documentation*, https://docs.python.org/3/library/csv.html. Accessed 18 Apr. 2025.

Das, Suman. “Fine Tune Large Language Model (LLM) on a Custom Dataset with QLoRA.” *Medium*, 25 Jan. 2024, https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07.

Data Collator. https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/data_collator. Accessed 25 Apr. 2025.

Datasets. https://huggingface.co/docs/datasets/index. Accessed 22 Apr. 2025.

Deepseek-R1. https://ollama.com/deepseek-r1. Accessed 4 Apr. 2025.

Defunct-Datasets/Eli5 · Datasets at Hugging Face. 6 Mar. 2025, https://huggingface.co/datasets/defunct-datasets/eli5.

"Different Ways to Create Pandas Dataframe." *GeeksforGeeks*, 00:25:10+00:00, https://www.geeksforgeeks.org/different-ways-to-create-pandas-dataframe/.

Different Ways to Create Pandas Dataframe | GeeksforGeeks. https://www.geeksforgeeks.org/different-ways-to-create-pandas-dataframe/. Accessed 18 Apr. 2025.

EASIEST Way to Fine-Tune a LLM and Use It With Ollama. 20 Sept. 2024, https://www.topview.ai/blog/detail/easiest-way-to-fine-tune-a-llm-and-use-it-with-ollama.

Embedding Models · Ollama Blog. https://ollama.com/public/Embedding models. Accessed 20 Apr. 2025.

Emerging Tech Services | Predictability Models Powered by WorldData.AI. https://worlddata.ai. Accessed 8 Apr. 2025.

Evaluate-Metric (Evaluate Metric). 10 Jan. 2025, https://huggingface.co/evaluate-metric.

Falbel, Daniel. *Posit AI Blog: Understanding LoRA with a Minimal Example*. June 2023. *blogs.rstudio.com*, https://blogs.rstudio.com/tensorflow/posts/2023-06-22-understanding-lora/.

Fine-Tuning. https://huggingface.co/docs/transformers/v4.51.3/en/training. Accessed 25 Apr. 2025.

Fine-Tuning a Vision Language Model (Qwen2-VL-7B) with the Hugging Face Ecosystem (TRL) - Hugging Face Open-Source AI Cookbook. https://huggingface.co/learn/cookbook/en/fine_tuning_vlm_trl#6-compare-fine-tuned-model-vs-base-model--prompting-. Accessed 25 Apr. 2025.

Fine-Tuning LLMs: A Guide With Examples. https://www.datacamp.com/tutorial/fine-tuning-large-language-models. Accessed 22 Apr. 2025.

Fine-Tuning Models with Ollama: A Comprehensive Guide. https://www.arsturn.com/blog/deep-dive-fine-tuning-models-ollama. Accessed 20 Apr. 2025.

Fine-Tuning With Ollama Techniques | Restackio. https://www.restack.io/p/fine-tuning-answer-ollama-techniques-cat-ai. Accessed 20 Apr. 2025.

Geiger, Stuart. *Staeiou/Gigo_qss_2021*. 2021. 5 July 2021. *GitHub*, https://github.com/staeiou/gigo_qss_2021.

Gemma3. https://ollama.com/gemma3. Accessed 4 Apr. 2025.

Hannun, Awni, et al. *Mlx*. 2023. 22 Apr. 2025. *GitHub*, https://github.com/ml-explore.

Hugging Face - Documentation. https://huggingface.co/docs. Accessed 25 Apr. 2025.

Huggingface/Evaluate. 2022. Hugging Face, 23 Apr. 2025. *GitHub*, https://github.com/huggingface/evaluate.

Huihui_ai/Qwq-Abliterated. https://ollama.com/huihui_ai/qwq-abliterated. Accessed 4 Apr. 2025.

Hyperparameter Search. https://huggingface.co/docs/transformers/hpo_train. Accessed 22 Apr. 2025.

JSON Lines. https://jsonlines.org/. Accessed 22 Apr. 2025.

Khushwant04/Research-Papers · Datasets at Hugging Face. https://huggingface.co/datasets/khushwant04/Research-Papers. Accessed 17 Apr. 2025.

Laion/COREX-18text · Datasets at Hugging Face. https://huggingface.co/datasets/laion/COREX-18text. Accessed 8 Apr. 2025.

Laion/Openalex-Metadata · Datasets at Hugging Face. https://huggingface.co/datasets/laion/openalex-metadata. Accessed 8 Apr. 2025.

Lan, Haoyong. *CMU LibGuides: Artificial Intelligence Research: Find Datasets*. https://guides.library.cmu.edu/artificial-intelligence/datasets. Accessed 8 Apr. 2025.

Llama3.3. https://ollama.com/llama3.3. Accessed 6 Apr. 2025.

Llama3.3/License. https://ollama.com/llama3.3/blobs/bc371a43ce90. Accessed 6 Apr. 2025.

Logging. https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/logging#logging. Accessed 23 Apr. 2025.

Low Rank Adaptation: A Technical Deep Dive. https://www.ml6.eu/blogpost/low-rank-adaptation-a-technical-deep-dive. Accessed 21 Apr. 2025.

“Mlx-Examples/Lora/Lora.Py at Main · Ml-Explore/Mlx-Examples.” *GitHub*, https://github.com/ml-explore/mlx-examples/blob/main/lora/lora.py. Accessed 22 Apr. 2025.

Moi, Anthony, and Nicolas Patry. *HuggingFace’s Tokenizers*. 2019. 0.13.4, Apr. 2023. *GitHub*, https://github.com/huggingface/tokenizers.

Non-Engineers Guide: Train a LLaMA 2 Chatbot. 30 Jan. 2025, https://huggingface.co/blog/Llama2-for-non-engineers.

Nous-Hermes2. https://ollama.com/nous-hermes2. Accessed 4 Apr. 2025.

“Ollama/Docs at Main · Ollama/Ollama.” *GitHub*, https://github.com/ollama/ollama/tree/main/docs. Accessed 17 Apr. 2025.

“Ollama/Docs/Api.Md at Main · Ollama/Ollama.” *GitHub*, https://github.com/ollama/ollama/blob/main/docs/api.md. Accessed 18 Apr. 2025.

Ollama/Docs/Template.Md at 1d99451ad705478c0a22262ad38b5a403b61c291 · Ollama/Ollama. https://github.com/ollama/ollama/blob/1d99451ad705478c0a22262ad38b5a403b61c291/docs/template.md?plain=1#L29. Accessed 18 Apr. 2025.

OllamaEmbeddings | 🦜🔗 LangChain. https://python.langchain.com/docs/integrations/text_embedding/ollama/. Accessed 17 Apr. 2025.

Ollama/Ollama-Python. 2023. Ollama, 20 Apr. 2025. *GitHub*, https://github.com/ollama/ollama-python.

Ollama/Ollama-Python: Ollama Python Library. https://github.com/ollama/ollama-python. Accessed 18 Apr. 2025.

Ollama/README.Md at Main · Ollama/Ollama. https://github.com/ollama/ollama/blob/main/README.md#quickstart. Accessed 17 Apr. 2025.

Open LLM Leaderboard - a Hugging Face Space by Open-Llm-Leaderboard. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard. Accessed 4 Apr. 2025.

Openthinker. https://ollama.com/openthinker. Accessed 4 Apr. 2025.

Open-Thoughts/Open-Thoughts. 2025. open-thoughts, 4 Apr. 2025. *GitHub*, https://github.com/open-thoughts/open-thoughts.

Open-Thoughts/OpenThoughts-114k · Datasets at Hugging Face. 11 Feb. 2025, https://huggingface.co/datasets/open-thoughts/OpenThoughts-114k.

Optimizers. https://huggingface.co/docs/transformers/v4.51.3/en/optimizers. Accessed 25 Apr. 2025.

Optimizing Your LLM in Production. https://huggingface.co/blog/optimize-llm. Accessed 23 Apr. 2025.

"Os — Miscellaneous Operating System Interfaces." *Python Documentation*, https://docs.python.org/3/library/os.html. Accessed 19 Apr. 2025.

Outperforming DeepSeekR1-32B with OpenThinker2. 3 Apr. 2025, https://portfolio-blog-starter.vercel.app/blog/thinkagain.

Overview. https://huggingface.co/docs/transformers/quantization/overview. Accessed 23 Apr. 2025.

Papers with Code - About Papers With Code. https://paperswithcode.com/about. Accessed 8 Apr. 2025.

Papers with Code - PWC Dataset Licensing Guide. https://paperswithcode.com/datasets/license. Accessed 8 Apr. 2025.

Part 2: Building Your Training Data for Fine-Tuning – Andy Peatling. https://apeatling.com/articles/part-2-building-your-training-data-for-fine-tuning/. Accessed 20 Apr. 2025.

---. 8 Jan. 2024, https://apeatling.com/articles/part-2-building-your-training-data-for-fine-tuning/.

---. 8 Jan. 2024, https://apeatling.com/articles/part-2-building-your-training-data-for-fine-tuning/.

Part 3: Fine-Tuning Your LLM Using the MLX Framework – Andy Peatling. 8 Jan. 2024, https://apeatling.com/articles/part-3-fine-tuning-your-llm-using-the-mlx-framework/.

PEFT. https://huggingface.co/docs/transformers/peft. Accessed 22 Apr. 2025.

Perplexity of Fixed-Length Models. https://huggingface.co/docs/transformers/perplexity. Accessed 25 Apr. 2025.

Phi4. https://ollama.com/phi4. Accessed 4 Apr. 2025.

Pietrusky, Stefan. “How to Talk to a PDF File Without Using Proprietary Models: CLI + Streamlit + Ollama.” *Towards Data Science*, 14 Aug. 2024, https://towardsdatascience.com/how-to-talk-to-a-pdf-file-without-using-proprietary-models-cli-streamlit-ollama-6c22437ed932/.

Pilone, Vinny. *Research Approach for Final Project*. 21 Mar. 2025, p. 1, https://docs.google.com/document/d/1V_7wq0KZYU0VZdKYYSamiaTZw-KLizVTnV3Nz7si5mk/edit?tab=t.0.

Pipelines. https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/pipelines#transformers.QuestionAnsweringPipeline. Accessed 25 Apr. 2025.

"Proper Way to Train Model on My Data and Load into Ollama? · Issue #7755 · Ollama/Ollama." *GitHub*, https://github.com/ollama/ollama/issues/7755. Accessed 17 Apr. 2025.

"Python - List Files in a Directory." *GeeksforGeeks*, 16:47:27+00:00, https://www.geeksforgeeks.org/python-list-files-in-a-directory/.

Question Answering. https://huggingface.co/docs/transformers/tasks/question_answering. Accessed 25 Apr. 2025.

Quickstart. https://huggingface.co/docs/transformers/v4.51.3/en/quicktour. Accessed 23 Apr. 2025.

Qwen2.5 - a Qwen Collection. 26 Feb. 2025, https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e.

QwenLM/Qwen: The Official Repo of Qwen (通义千问) Chat & Pretrained Large Language Model Proposed by Alibaba Cloud. https://github.com/QwenLM/Qwen. Accessed 6 Apr. 2025.

QwenLM/QwQ. 2025. Qwen, 6 Apr. 2025. *GitHub*, https://github.com/QwenLM/QwQ.

QwenLM/QwQ: QwQ Is the Reasoning Model Series Developed by Qwen Team, Alibaba Cloud. https://github.com/QwenLM/QwQ. Accessed 6 Apr. 2025.

Qwen/Qwen2.5-1.5B · Hugging Face. 26 Feb. 2025, https://huggingface.co/Qwen/Qwen2.5-1.5B.

Qwen/Qwen2.5-14B · Hugging Face. 26 Feb. 2025, https://huggingface.co/Qwen/Qwen2.5-14B.

Qwen/QwQ-32B · Hugging Face. 6 Mar. 2025, https://huggingface.co/Qwen/QwQ-32B.

Qwq. https://ollama.com/qwq. Accessed 4 Apr. 2025.

Rexarski/Eli5_category · Datasets at Hugging Face. 17 Dec. 2024, https://huggingface.co/datasets/rexarski/eli5_category.

Shrivastav, Shivang. "Demystifying the Advantage Function in Reinforcement Learning." *Medium*, 29 Dec. 2024, https://shivang-ahd.medium.com/demystifying-the-advantage-function-in-reinforcement-learning-1c2b2a0d0daa.

Skymind AI - Enterprise Platform | Profile. https://yippy.com/yp/skymind. Accessed 8 Apr. 2025.

Starcoder2. https://ollama.com/starcoder2. Accessed 4 Apr. 2025.

Tasks - Hugging Face. https://huggingface.co/tasks. Accessed 23 Apr. 2025.

Team, Qwen. "About Us." *Qwen*, https://qwenlm.github.io/about/. Accessed 6 Apr. 2025.

---. "QwQ-32B: Embracing the Power of Reinforcement Learning." *Qwen*, 6 Mar. 2025, https://qwenlm.github.io/blog/qwq-32b/.

"The MIT License." *Open Source Initiative*, https://opensource.org/license/MIT. Accessed 8 Apr. 2025.

Tianyi. *Tiiiger/Bert_score*. 2019. 22 Apr. 2025. *GitHub*, https://github.com/Tiiiger/bert_score.

"Time.Sleep() in Python." *GeeksforGeeks*, 00:10:51+00:00, https://www.geeksforgeeks.org/sleep-in-python/.

Tokenizer. https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/tokenizer. Accessed 23 Apr. 2025.

Tokenizers. https://huggingface.co/docs/transformers/fast_tokenizers. Accessed 23 Apr. 2025.

Torch.Compile. https://huggingface.co/docs/transformers/perf_torch_compile. Accessed 23 Apr. 2025.

Trainer. https://huggingface.co/docs/transformers/v4.51.3/en/main_classes/trainer. Accessed 25 Apr. 2025.

Transformers. https://huggingface.co/docs/transformers/index. Accessed 22 Apr. 2025.

"Transformers/Src/Transformers/Trainer.Py at 052e652d6d53c2b26ffde87e039b723949a53493 · Huggingface/Transformers." *GitHub*, https://github.com/huggingface/transformers/blob/052e652d6d53c2b26ffde87e039b723949a53493/src/transformers/trainer.py. Accessed 22 Apr. 2025.

"Tuning in 5, 15, 50 Minutes." *Project Name Not Set*, https://ravinkumar.com/GenAiGuidebook/deepdive/deepdive/SmallModelFinetuning.html. Accessed 20 Apr. 2025.

Tuning in 5, 15, 50 Minutes — The GenAI Guidebook. https://ravinkumar.com/GenAiGuidebook/deepdive/SmallModelFinetuning.html. Accessed 20 Apr. 2025.

Tutorial: How to Finetune Llama-3 and Use In Ollama | Unsloth Documentation. 16 Feb. 2025, https://docs.unsloth.ai/basics/tutorial-how-to-finetune-llama-3-and-use-in-ollama.

UCI Machine Learning Repository. https://archive.ics.uci.edu/. Accessed 8 Apr. 2025.

Unslothai/Unsloth. 2023. Unsloth AI, 25 Apr. 2025. *GitHub*, https://github.com/unslothai/unsloth.

"Venv — Creation of Virtual Environments." *Python Documentation*, https://docs.python.org/3/library/venv.html. Accessed 25 Apr. 2025.

"Weights & Biases." *W&B*, httpss://wandb.ai/byyoung3/mlnews2/reports/Fine-Tuning-Llama-3-with-LoRA-TorchTune-vs-HuggingFace--Vmlldzo3NjE3NzAz. Accessed 22 Apr. 2025.

What Is Text Generation? - Hugging Face. 27 Aug. 2024, https://huggingface.co/tasks/text-generation.

Younger: The First Dataset for Artificial Intelligence-Generated Neural Network Architecture | AI Research Paper Details. https://aimodels.fyi. Accessed 17 Apr. 2025.

Yu, Xiaojian. "Fine-Tuning Llama3.1 and Deploy to Ollama." *Medium*, 2 Sept. 2024, https://medium.com/@yuxiaojian/fine-tuning-llama3-1-and-deploy-to-ollama-f500a6579090.

---. "Fine-Tuning Ollama Models with Unsloth." *Medium*, 22 Aug. 2024, https://medium.com/@yuxiaojian/fine-tuning-ollama-models-with-unsloth-a504ff9e8002.

Yugank .Aman. "Top 10 Open-Source LLM Models and Their Uses." *Medium*, 3 Feb. 2025, https://medium.com/@yugank.aman/top-10-open-source-llm-models-and-their-uses-6f4a9aced6af.