**Foundational LLM Concepts**

[**Intro to Large Language Models**](https://www.youtube.com/watch?v=zjkBMFhNj_g)

A general-audience overview of how LLMs are trained and the key concepts behind their operation, including pre-training, fine-tuning, and RLHF.

[**Let's build GPT: from scratch, in code, spelled out**](https://www.youtube.com/watch?si=DdtaubkIDlafzAQV&v=kCc8FmEb1nY&feature=youtu.be)

A deep, code-first dive into the Transformer architecture that powers models like GPT.

[**LLM Course on GitHub**](https://github.com/mlabonne/llm-course)

A comprehensive course covering LLM fundamentals, the science of building LLMs, and the engineering of LLM-based applications.

**Library-Specific Tutorials**

[**RDKit**](https://www.rdkit.org/docs/GettingStartedInPython.html)

RDKit is a collection of cheminformatics and machine-learning software written in C++ and Python.

[**PySCF**](https://pyscf.org/)

The Python-based Simulations of Chemistry Framework (PySCF) is an open-source library for quantum chemistry calculations. It is highly extensible and designed for simplicity, both for users and developers.

[**Atomic Simulation Environment (ASE)**](https://ase-lib.org/tutorials/tutorials.html)

The Atomic Simulation Environment (ASE) is a set of tools and Python modules for setting up, manipulating, running, visualizing and analyzing atomistic simulations.

[**Pymatgen**](https://pymatgen.org/pymatgen.html)

Pymatgen (Python Materials Genomics) is a robust, open-source Python library for materials analysis.

[**LangChain**](https://python.langchain.com/docs/introduction/)

LangChain is an open-source library specifically designed for creating applications using large language models (LLMs).

[**LangGraph**](https://langchain-ai.github.io/langgraph/)

LangGraph, created by LangChain, is an open source AI agent framework designed to build, deploy and manage complex generative AI agent workflows.

**Technique-Specific Guides**

[**Fine-Tuning LLMs**](https://huggingface.co/docs/transformers/training)

Fine-tuning allows you to adapt pre-trained models to specific tasks or domains, making them more effective for specialized applications in materials science and chemistry.

[**Retrieval-Augmented Generation (RAG)**](https://python.langchain.com/docs/tutorials/rag/)

RAG combines the power of retrieval systems with language generation, allowing LLMs to access and use external knowledge bases effectively.

[**Prompt Engineering**](https://www.promptingguide.ai/)

Learn how to craft effective prompts to get the best results from language models in scientific applications.

**Materials Science & Chemistry Datasets**

[**Materials Science Datasets Compilation**](https://github.com/blaiszik/awesome-matchem-datasets)

A curated list of awesome materials and chemistry datasets by [**Ben Blaiszik**](https://scholar.google.com/citations?user=J-x5n7IAAAAJ&hl=en).

**General Datasets**

[**arXiv Preprints**](https://arxiv.org/)

The entire collection of preprints from arXiv is available for bulk download, providing a massive corpus for text mining the latest scientific research.

[**Hugging Face Datasets**](https://huggingface.co/datasets)

A central hub for thousands of datasets, including a growing number for materials science and chemistry.

[**Kaggle Datasets**](https://www.kaggle.com/datasets)

A platform hosting a wide variety of public datasets, including many relevant to chemistry and materials science.