# Literature and Resources

Good books and resources to read

This section provides a curated list of books and resources to enhance your understanding of Large Language Models. Each recommendation includes a brief description to help you choose the most suitable resources for you.

# Large Language Models

## Hosted LLMs

These are powerful LLMs hosted by companies, which you can access through APIs (Application Programming Interfaces). You typically pay for usage.

- OpenAI (ChatGPT): The creators of ChatGPT and GPT-4, offering a range of models.
- Mistral: A European-based company offering competitive models.
- Google (Gemini): Google's LLM, offering strong performance and integration with Google services.
- Anthropic (Claude): Known for its its ability to handle code effectively.

## Local LLMs

These tools can be used to run open-source LLMs that you can download and run on your own machine. This gives you more privacy and control, but requires more technical expertise and computational resources.

- Ollama: Free and open-source tool to run large language models locally, supports a wide range of models. Note, that the models are not as powerful as the hosted ones and that your computer needs to have a good GPU to run larger models. Smaller models with less than 8B parameters can also often be run on a CPU with enough available RAM. Great for privacy and if you don't want to pay for the hosted models.
- Hugging Face: Hosts a wide range of large language models, including models fine-tuned for specific tasks by the community. Models can also be downloaded to Ollama and run locally, if your computer is powerful enough.

## Working with data

In addition to the hosted and local LLMs, there are also tools that allow you to work with LLMs in a browser to build RAG apps or custom chatbots.

- NotebookLM**: Google's Gemini that can be fed with files, images and YouTube videos to generate text based on the content. Only works within a workspace of Google, you can't make it available to the public (yet).
- LM Studio: An application for discovering, downloading, and running LLMs locally. It supports various model architectures and offers both a Chat UI and an OpenAI-compatible local server. Features include offline document chat capabilities and easy model downloads from Hugging Face.
- Open Web UI**: Open Web UI is a tool to run large language models locally (in conjuction with, for example, Ollama). It is a browser-based interface that allows you to interact with the models and build RAG apps.

## Further resources

- Quarto
  - A static website generator that is very powerful and flexible. Used to create the slides and the website for the course.
- Github
  - The largest provider for git repositories owned by Microsoft. A lot of open source projects are hosted here and you can read the code.
- Daily Dose of Data Science
  - A website and a newsletter with lots of easy-to-digest resources to improve your skills in Data Science and Large Language Models.