Mathematics of LLMs:

Which models and technologies exist other than ChatGPT and what's different/special about them?

There are a multitude of technologies that are LLMs each with their own specialised uses.
A few I have found particularly useful are:

- Gemini 2.5 is great for a large context window and analysing large documents as it can process up to 1 million tokens with gpt 4 only handling a maximum of 32,768 in its context window.

- I also looked at notebook LM which also has a large context window so you can put whole articles and it can generate reports and also things like podcast snapshots which are incredibly helpful for learning.

- OLLAMA is a way to run LLMs locally. This has many advantages such as privacy most LLMs such as chatGPT use or at least have access to all data that is input into it (unless you pay for the privilege). If you run an LLM locally there is no threat that sensitive data can be stolen so it can be utilised fully. However you will need a good amount of compute to be able to run it. After learning about fine tuning I realised that you can use OLLAMA for fine tuning so you could run a specialised model on your computer.

- I came across Loveable which specializes in developing software. It seems to hold up better than ChatGPT whereby it checks its errors as it's going more thoroughly leading to a much better result. It also does security checks to make it much more production friendly.

Learn and discuss some terminology. I made some suggestions, but you are encouraged to add some of your own. Don't just learn a definition, be prepared to talk about it in some detail:

I have researched this terminology:

Slop - describes low to mid tier AI generated content that isn't really useful such as Memes or poorly written blogs/articles -

Hallucinations - it describes when an LLM outputs a result as if it were true but is actually false

Context Window - The amount of text and data an AI model can consider at any one time to process information and generate a response

Fine Tuning - the process of taking a pre-trained AI model and training it further on a smaller, more specific dataset to adapt it to a new or specialized task.

Token - Tokens are tiny units of data that come from breaking down bigger chunks of information. - this is also the currency of LLMs denoting the cost of requiring outputs

Encoder - essential for a machine learning model that transforms raw input data into a compact, numerical representation which is a type of vector.

Bias - In AI bias describes outputs that are affected by pre-training such as opinions of programmers such as what they believe is right and wrong or whether something is useful or not. Or simply the information it is trained upon.

Zero shot learning - an AI technique where a model can identify and classify categories it has never seen during training.

Spend some time reading the literature, watching YouTube videos and working through tutorials. Add to the document which ones you found particularly useful and why.

1. https://www.youtube.com/watch?v=wjZofJX0v4M&t=528s this is a video by 3blue1brown called Transformers, the tech behind LLMs. - 3blue1brown uses very intuitive diagrams. It makes a more intermediate level of explaining the science and maths behind transformers more understandable.
2. https://www.youtube.com/watch?v=59bMh59JQDo&t=42s this is a video by google that states the 3 types of bias behind LLMs - This is a concept that really intrigues me as it has such large real world effects especially as these tools are such large parts of how are economies work.
3. https://medium.com/@amallya0523/how-an-llm-understands-input-the-math-under-the-hood-114ac69f96c6 - an article on the mathematics behind LLMs - Goes a bit more in depth on encoders and decoders
4. https://www.ibm.com/think/topics/zero-shot-learning - an article by IBM on zero shot learning - after researching some terminology on the topic i came across Zero shot learning this IBM article is very informative on it and has helped me gain a better grasp on the concept.
5. https://www.youtube.com/watch?v=_zfN9wnPvU0&t=482s - A kurzgesagt video on AI slop they work on the case of how their own research is inaccurate due to using certain research papers or articles that are wrong as they are AI generated or that there is so much Slop that sometimes AI will use AI slop in its own answers as facts creating a spiral of misinformation presented as facts.
6. https://learn.deeplearning.ai/courses/finetuning-large-language-models/lesson/ig7ql/why-finetune - I did a 1 hour fine tuning course, I learnt about how to fine tune a model and gained an insight in how business may use this technique to gain more specialised solutions.