Making sense of LLMs :

WebLLMs : [Web LLM : Bring LLM chatbots to browsers](https://www.kdnuggets.com/2023/05/webllm-bring-llm-chatbots-browser.html)

* Discusses use of WebGPU to reduce reliance on infra

[Customize llm like gpt3.5 with your own data and documents:](https://bdtechtalks.com/2023/05/01/customize-chatgpt-llm-embeddings/)

* Makes use of embedding through APIs to call the capture the semantics of the dat a
* Use vector DBs to store the daa