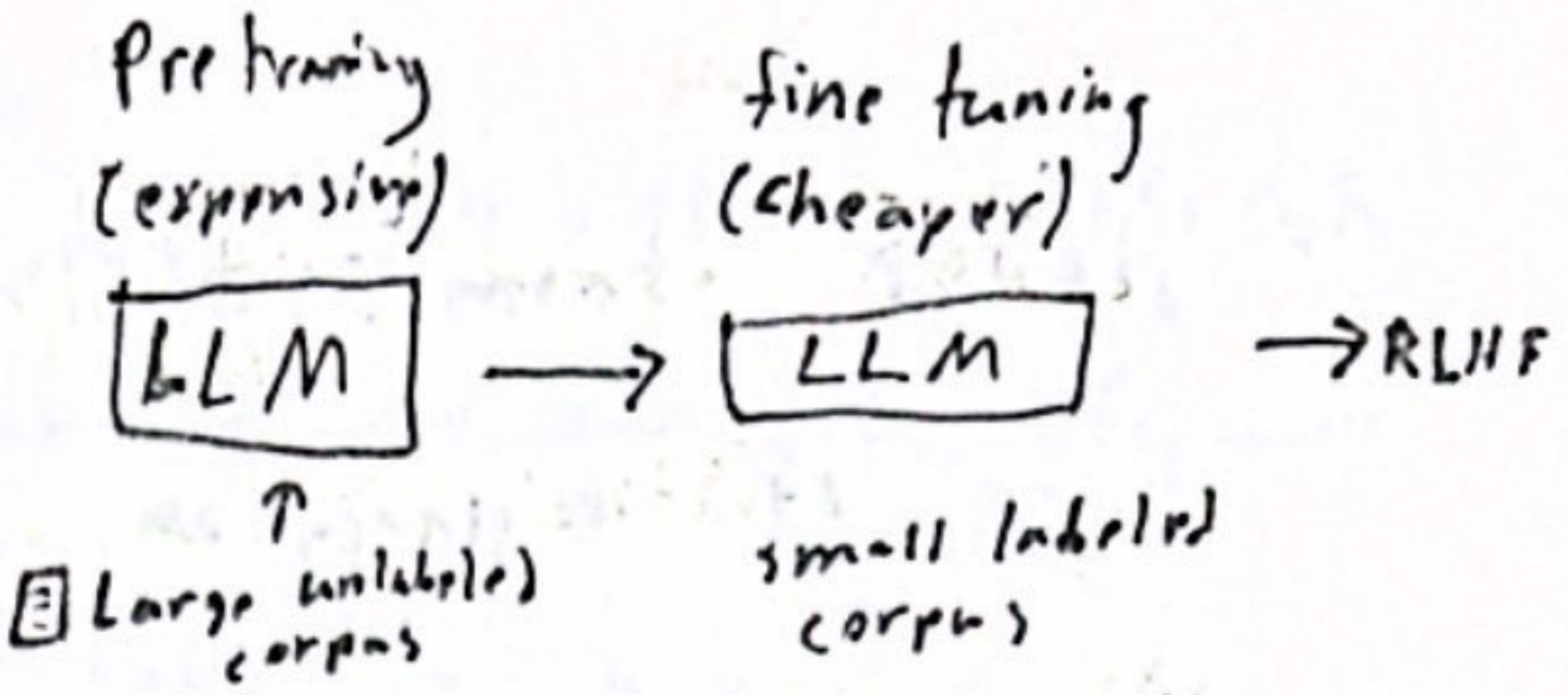
- Step 1 of the traning can also be devided into 2 parts

- Pretraning focuses on predicting next word and understanding language and context
- fine tuning is the model tuned on small specific tasks here the weights are adjusted for better outputs



- LLM's use GPU's because they preform alot of tasks in parrallel

- but not all LLM's are easily parralized. before 2017 most Language models processed text one word at a time

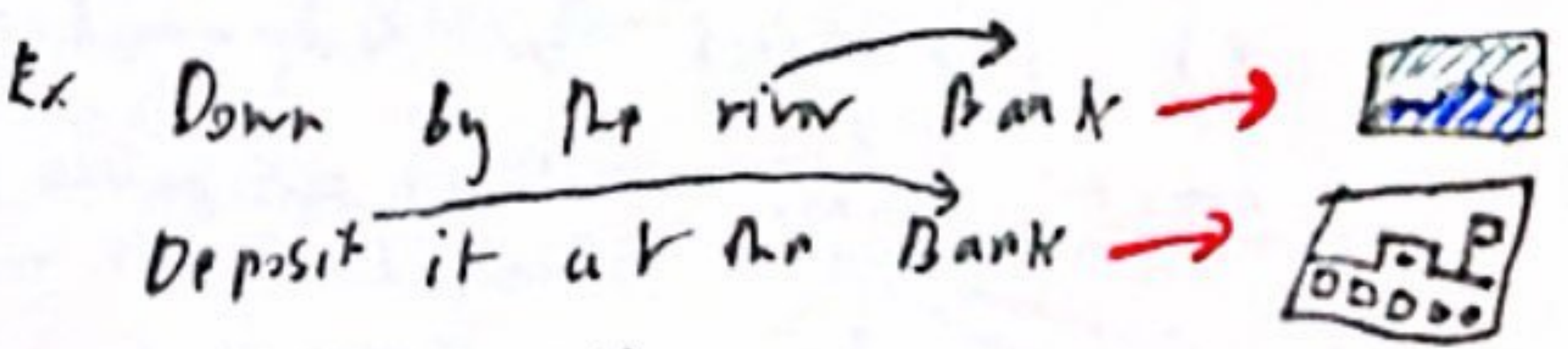
✱ but a team at google introduced a transformer they don't read text from the start to finish they soak it all in at once in parrallel

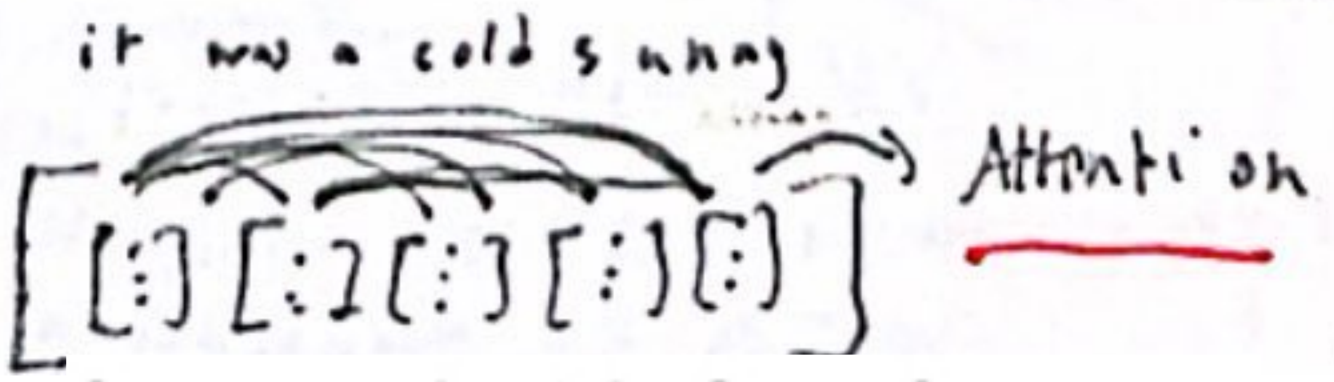
- The very first step inside a transformer and most LM's is to associate each word with a vector and make a long list of numbers as LLM's only work with continuous values ie numbers so we need to encode language as numbers

$$\left[\begin{array}{ccccc} \text{it} & \text{was} & \text{a} & \text{cold} & \text{sunny} \\ \downarrow & \downarrow & \downarrow & \downarrow & \\ \begin{bmatrix}1\\2\\3.4\\-5\end{bmatrix} & \begin{bmatrix}2\\5\\3.1\\-2.1\end{bmatrix} & \begin{bmatrix}1.1\\2.2\\-3\\-4\end{bmatrix} & \begin{bmatrix}1.2\\3.5\\2\\-1.1\end{bmatrix} & \begin{bmatrix}0\\-5.5\\-6\\-7\end{bmatrix}\end{array}\right] \quad ???$$

✱ we want this numbers ie vector to encode the meaning of the Corresponding word some how

- What makes transformers unique is there use of a special operation known as attention this operation gives these list on nums (word) a chance to talk to one another and refine the meaning they encode based on the context around all done in parallel for Ex in "Down by the river Bank" the word "Bank" can be changed based on the surronding context to encode the more specific notion of a river bank that a bank




it was a cold sunny
Attention