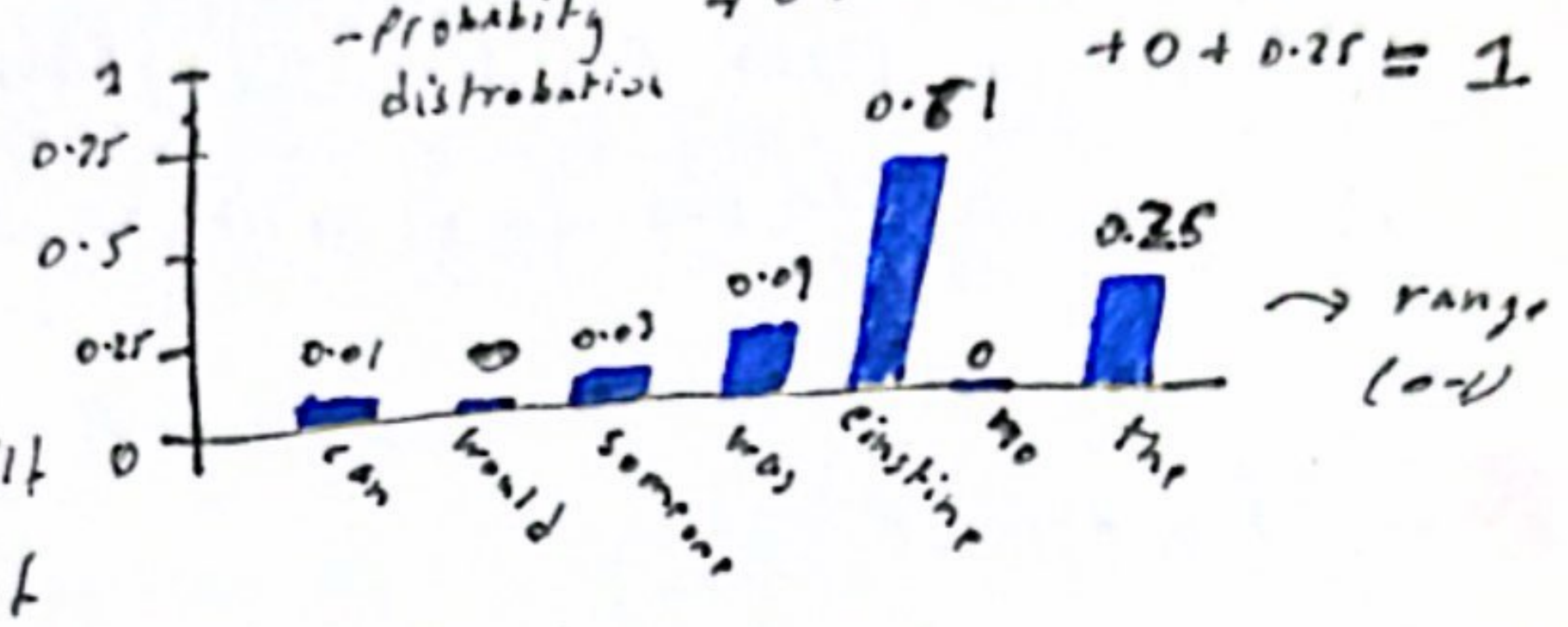
Softmax (LLM notes)

- if you want a sequence of numbers to act as a probability distribution ex a distribution over all next words, then each value has to be between 0 and 1 and for all of them to add up to 1

- but since we have alot of dot products and matrix multiplication the output we have by default dont follow these rules of probability distrabution

$\Rightarrow 0.01 + 0 + 0.03 + 0.09 + 0.61 + 0 + 0.25 = 1$



- Softmax is a way to turn a List of arbitrary numbers into a valid distrabution is such a way that the largest values end up closer to 1 and the smaller value end up closer to 0

- Logits input $x_1 - x_7$
- Logits output

$$\begin{bmatrix} -0.8 \\ -5.0 \\ +0.5 \\ +1.5 \\ +3.4 \\ -2.3 \\ +2.5 \end{bmatrix} \rightarrow \overset{\text{Softmax}}{\boxed{\begin{matrix} e^{x_1} / \sum_{n=1}^{N} e^{x_n} \\ \vdots \; x_1 - x_7 \\ e^{x_7} / \sum_{n=1}^{N} e^{x_n} \end{matrix}}} \rightarrow \begin{bmatrix} 0.01 \\ 0.00 \\ 0.03 \\ 0.09 \\ 0.61 \\ 0.00 \\ 0.25 \end{bmatrix}$$

$$\text{Softmax}(x)_i = \frac{e^{x_i}}{\sum_{n=1}^{N} e^{x_n}}$$

- this is known as a normalization step

- in this way if one input is meaningfully bigger (3.4) then it dominates the distribution and vise versa so if we were sampling from this we would almost alway pick the largest and most meaning full one (best prediction)

- Tempreture (Creativity)

- we can add a constant T to the denominator of Softmax

$$\text{Softmax with temp}(X)_i = \frac{e^{x_i}/T}{\sum^{N} e^{x_n}/T}$$

- if T is larger more weight is given to the Lower values boosting them up and vise versa at T=0 all weight =1 goes to the max input

$$T=9 \begin{bmatrix} 6 \\ -5 \\ 4 \\ 1.5 \\ 9.9 \\ -2.3 \\ 4 \end{bmatrix} \rightarrow \underset{\text{Temperture}}{\underset{\text{With}}{\text{Softmax}}} \rightarrow \begin{bmatrix} 0.18 \\ 0.06 \\ 0.15 \\ 0.11 \\ 0.27 \\ 0.08 \\ 0.15 \end{bmatrix}$$

$$T=2 \begin{bmatrix} 6 \\ -5 \\ 4 \\ 1.5 \\ 9.9 \\ -2.3 \\ 4 \end{bmatrix} \rightarrow \underset{\text{temperature}}{\underset{\text{with}}{\text{Softmax}}} \rightarrow \begin{bmatrix} 0.11 \\ 0 \\ 0.04 \\ 0.01 \\ 0.79 \\ 0 \\ 0.04 \end{bmatrix}$$