Normalization in LLMs

- Normalization in LLMs is typically applied between the attention and Feed forward layers (and vise versa) its usally applied before or after each sub layer (depending on variants):
  - post norm: normalization after Atention/MLP (used in original transformer paper)
  - pre norm; normalization before Attention/MLP (used in modern LLMs as it stabolizer training in deep models)

Why? :

- prevents exploding/vanishing activations: means the numbers (output of layers) dont become too big or small as they pass through many layers - this keeps the models computation stable
- Makes optimizations smoother: means the model learns more stedily during training, the loss surface becomes less chaotic, so gradient decent can take more reliable steps towords better preformance insted of bouncing around/getting stuck
- Allows deeper, more stable transformer training: beacuse norm keeps each layers well behaved (balanced and stable not too big or small, consistant across layers) you can stack more layers without the models training becoming unstable or diverging, in short normalization makes it possible for very deep NN. Like LLMs to still learn effectively and converge smoothly, insted of breaking down as they grow in size