Encoders and Decoders

- Encoders and decoders are two key components in many ML models, especially for sequence to sequence tasks like LLM
  - encoders take inputs (like a sentence, img or audio) and tranform it into a compact, informative repersentation that captures its meaning or features
  - the decoders then uses that repersentation to generate or reconstruct the desired output, such as a translated sentence, samarized text or reconstructed img.

Ex (in transformers): in transformers used for translation the encoder reads and processes the input sentence (eg in english) to produce contextual embeddings that capure the meaning of each word in relation to the whole sentence. The decoder than takes these embeddings and generates the output sentence (ie in French) one word at a time using attention to focus on the most relevent parts on the encoded input

Normalization and standerdization

- Normalization is the process of scalling data so all features are within a specific range typically [0,1] or [-1,1] this ensures that no feature dominates others simply because it has large numaric values
  - for min-max normalization (most common) $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$ → Scale [0,1]
  - for feature saling to [-1,1] do: $x_{norm} = 2 \times \frac{x - x_{min}}{x_{max} - x_{min}} - 1$ → do for each data point

- Standerdization (also called z-score norm) rescale data so that it has a mean (μ) of 0 and a standerd deviation (σ) of 1, this centers the data and ensures all features have comporiable variance formula: $z = \frac{x - \mu}{\sigma}$ → for each point

- Both together help model convergence/stabe training by ensuring features contribute equally

Rotary Positional Embeddings (RoPE)

- RoPE are a a way to encode token positions (positional embeddings see pg 112 for what those are) in transformers by roatating the query and key vectors in multi-headed attention, allowing the model to capture relitive positions efficiently without adding seperate position vectors, it works because the dot product of rotated K, Q vectors naturally encodes relitive position, why: • gen well to longer text sequences
  • is more efficient, cuts extra vectors