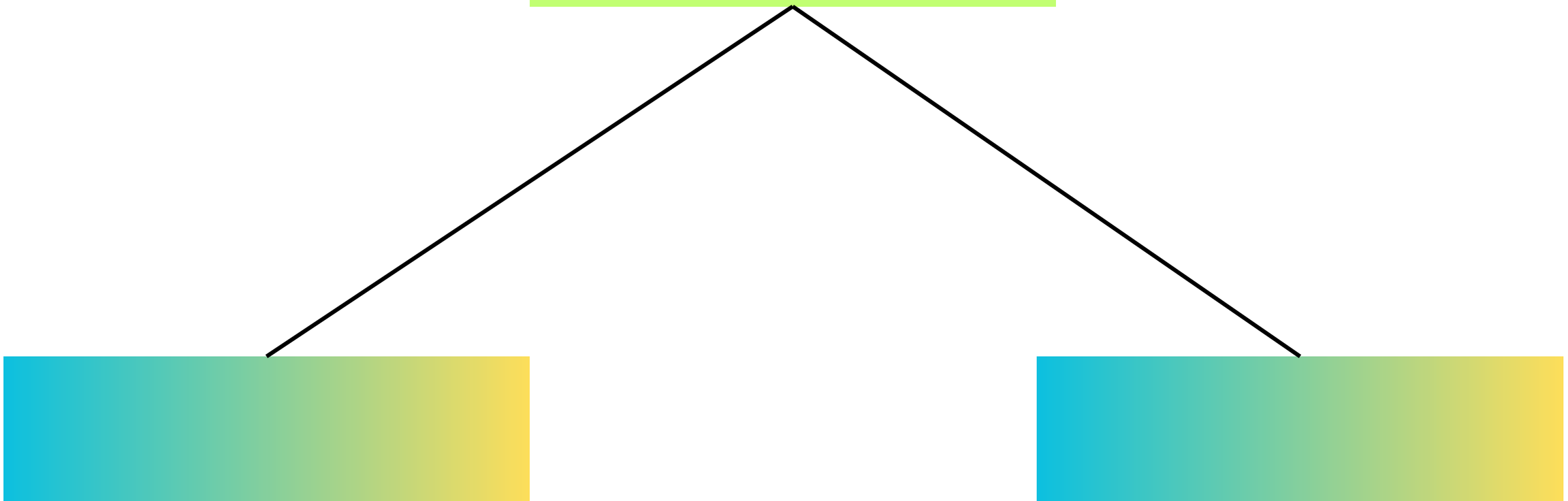
Transformers

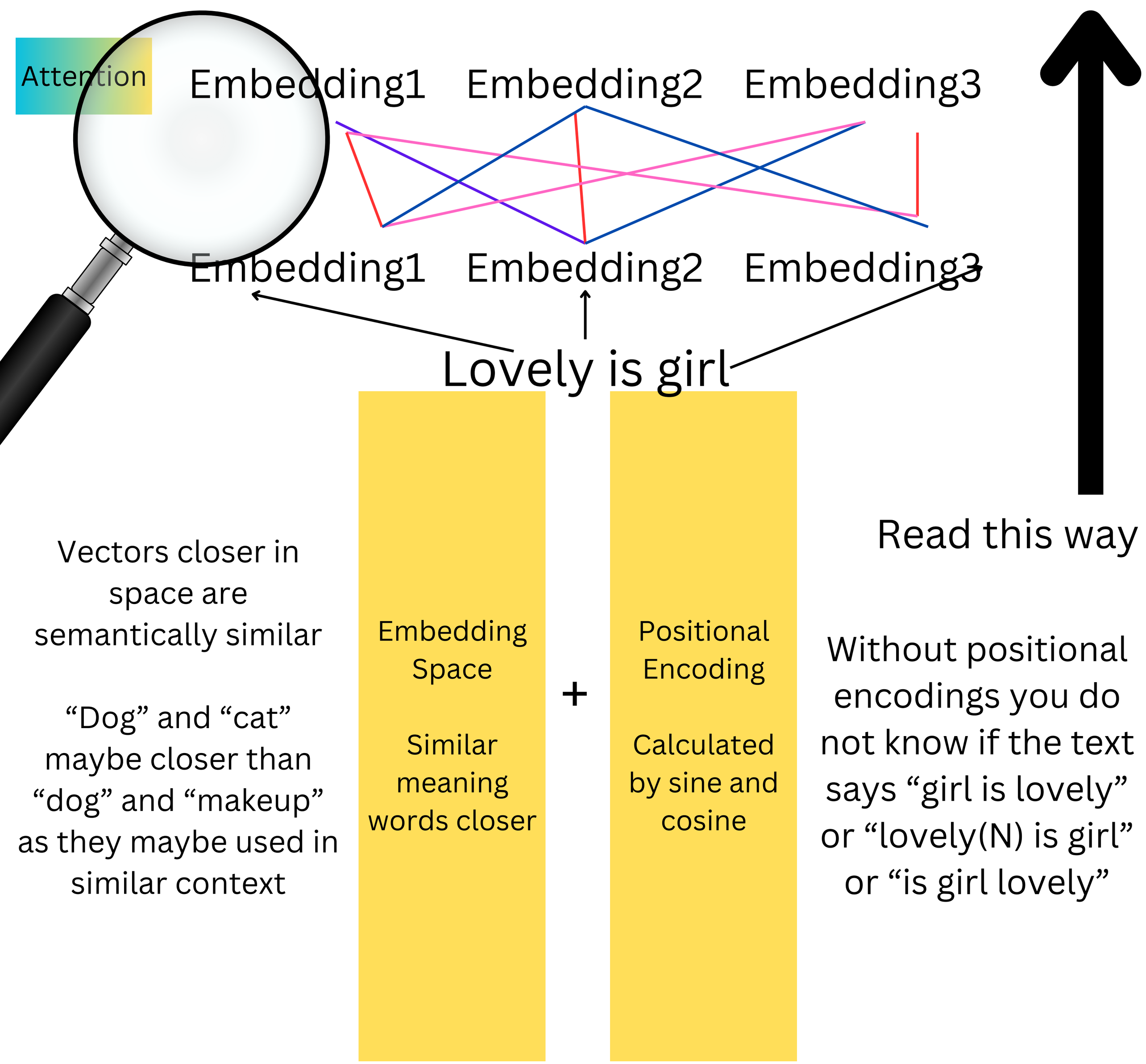


Encoders encode information and Decoders Decode it

The encoder:

Embedding Space + Attention Layers

Each attention layer captures a topic on what to pay attention to in their connections. I may pay attention to how close they are in embedding space with red being very close and blue being very far.
This can be done on multiple topics, and the model chooses on what it pays attention on. It may pay attention on color or gender based on the connections



Vectors closer in space are semantically similar

“Dog” and “cat” maybe closer than “dog” and “makeup” as they maybe used in similar context

Embedding Space

Similar meaning words closer

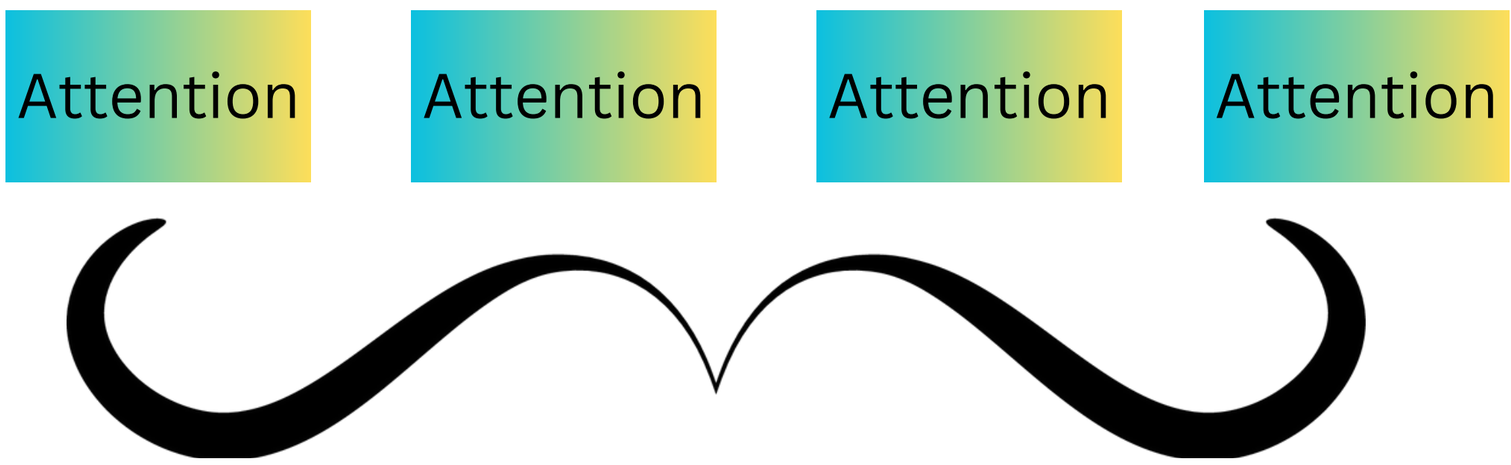
+

Positional Encoding

Calculated by sine and cosine

Without positional encodings you do not know if the text says “girl is lovely” or “lovely(N) is girl” or “is girl lovely”

Multi Headed Attention



Weighted addition of all the attention layers gives me all the information I need to know about the particular sentence in an embedding

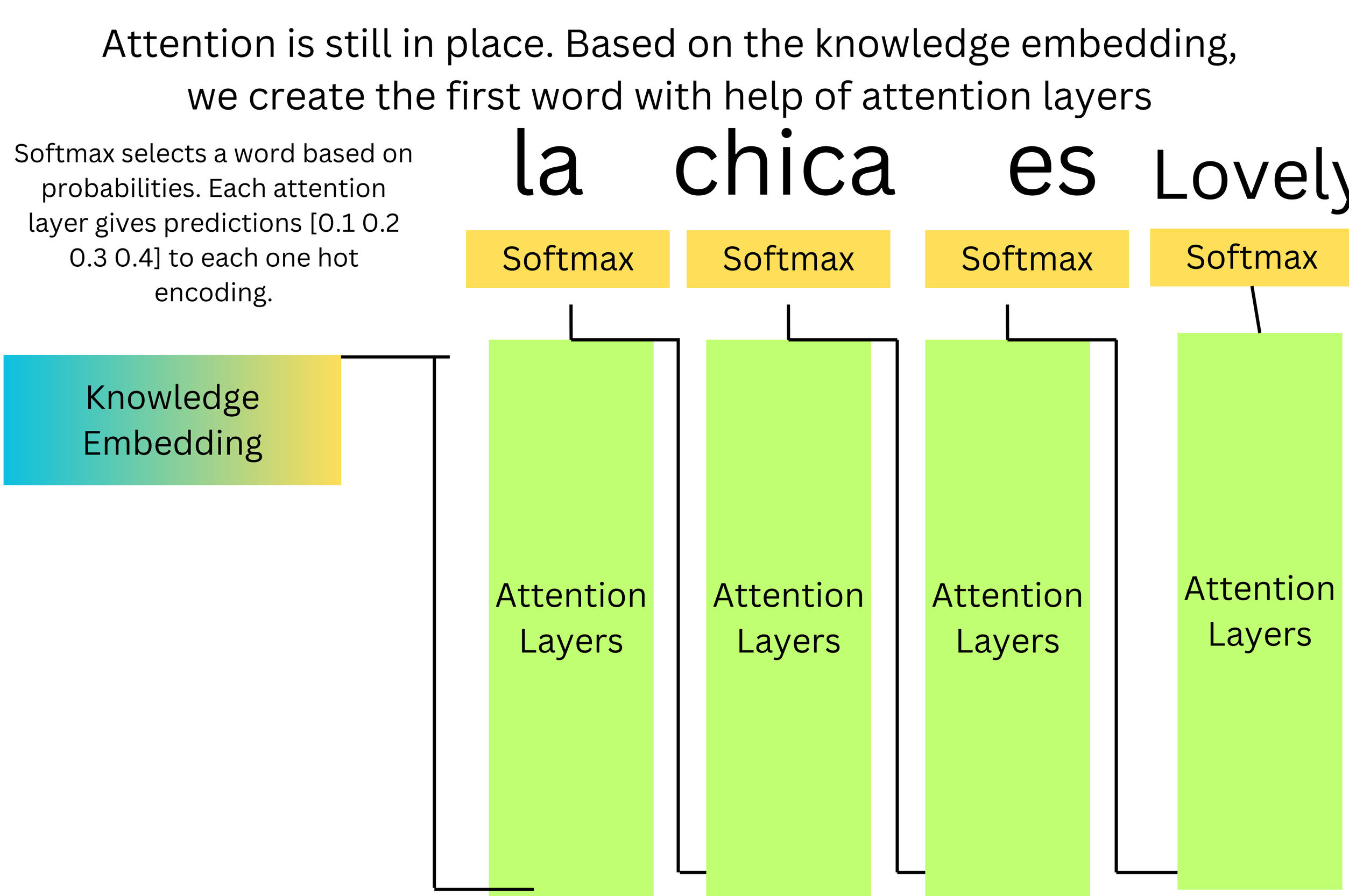
Knowledge Embedding

The Encoder

The Decoder:

Attention is still in place. Based on the knowledge embedding, we create the first word with help of attention layers

Softmax selects a word based on probabilities. Each attention layer gives predictions [0.1 0.2 0.3 0.4] to each one hot encoding.



Decoder-only Model: GPT
Encoder only: BERT families

Encoder+ Decoder: T5, BART, Transformer

GPT-4 has roughly 1.8 trillion parameters.

Hallucinations - It imagines something that isn’t real

Three parts:
Training on data
Post training : Fine tune the model to specific data. Finance data or Chemistry or Biology depending on context
Post training RL: Reinforce model understanding by giving model output.

Notable Mentions: Artificial General Intelligence and game of Chess