NLP4Web
Practice Session 9

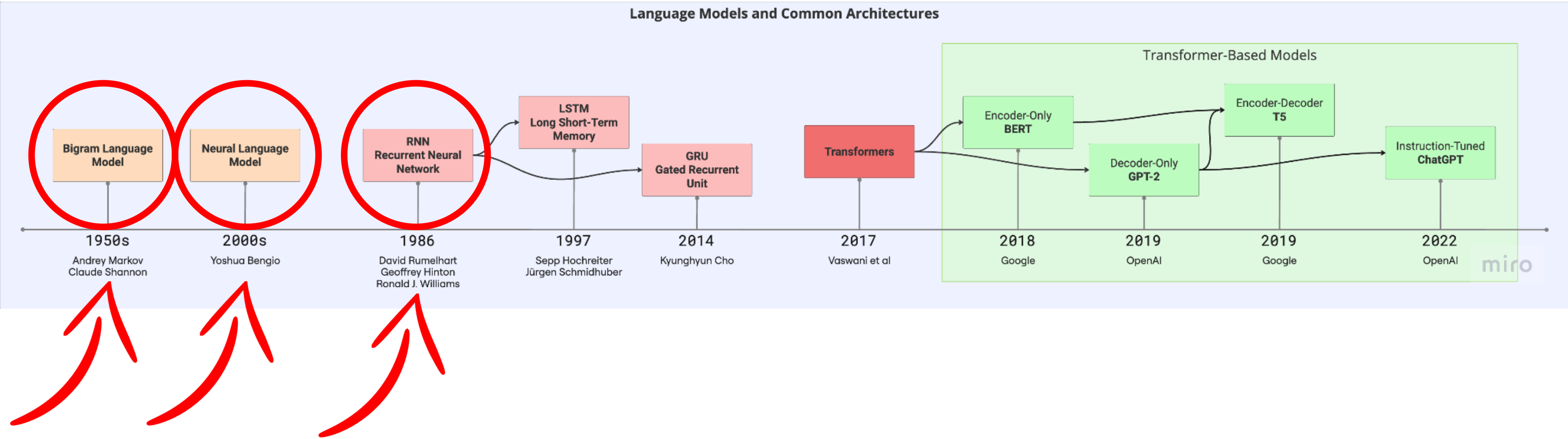
Transformers
Decoder-only (GPT)

Hovhannes Tamoyan

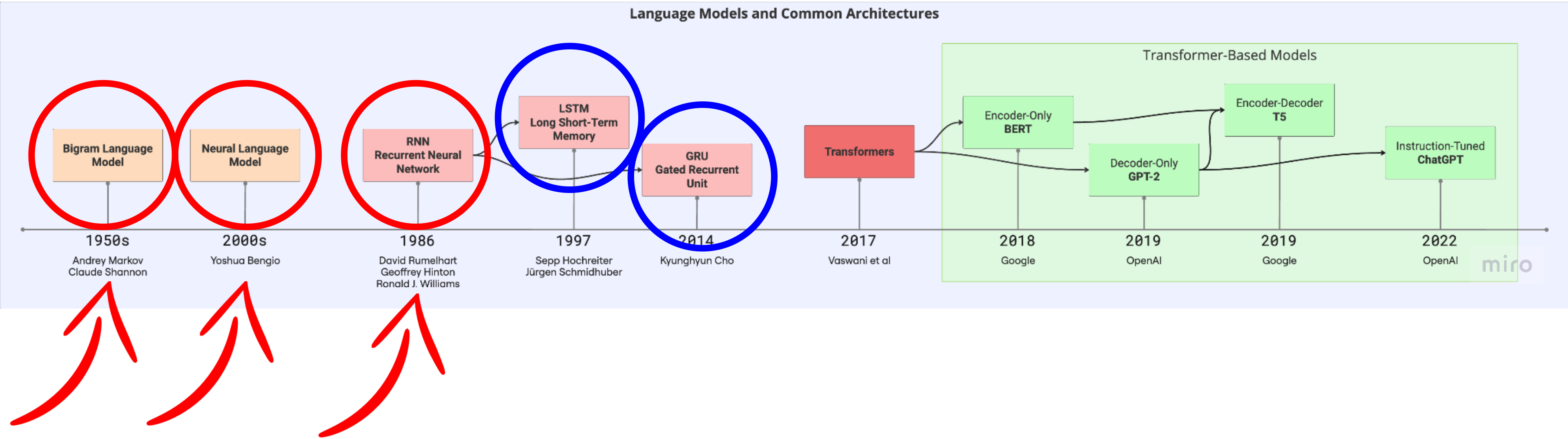
tamohannes.com

To not get lost in space over time, let’s
Use a **mind map**

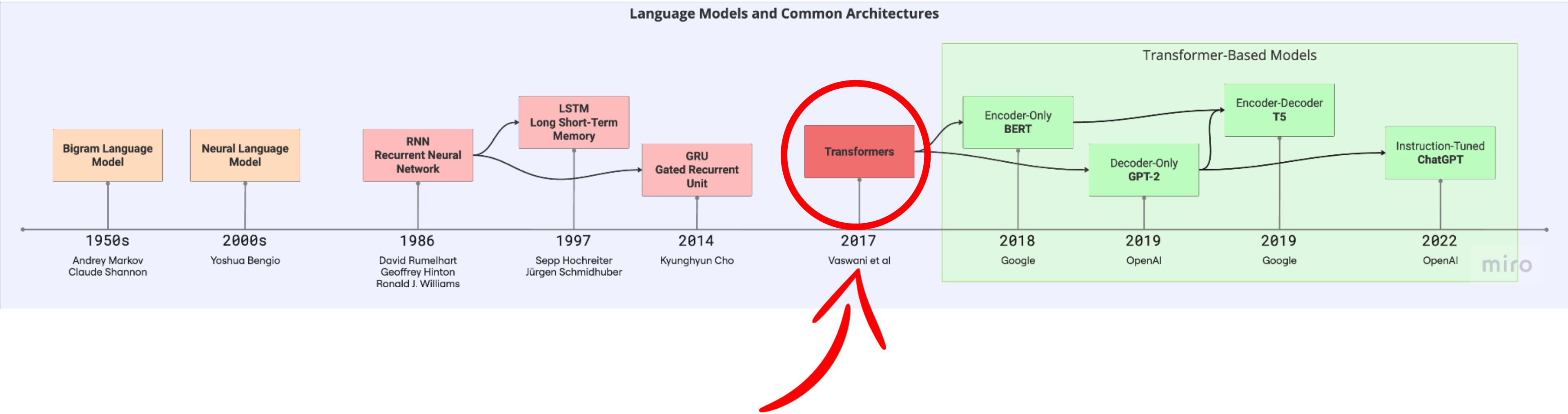
Last time we covered: Bigram LM, NLM, RNN



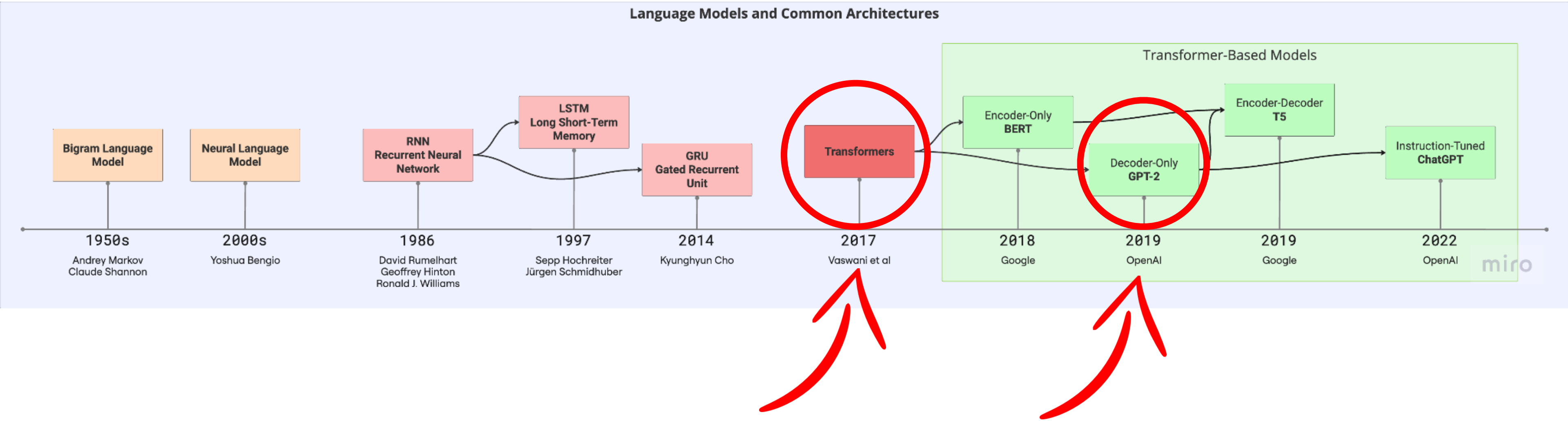
LSTM and GRU are for HW4



Today’s subject: Transformers (Decoder-only)

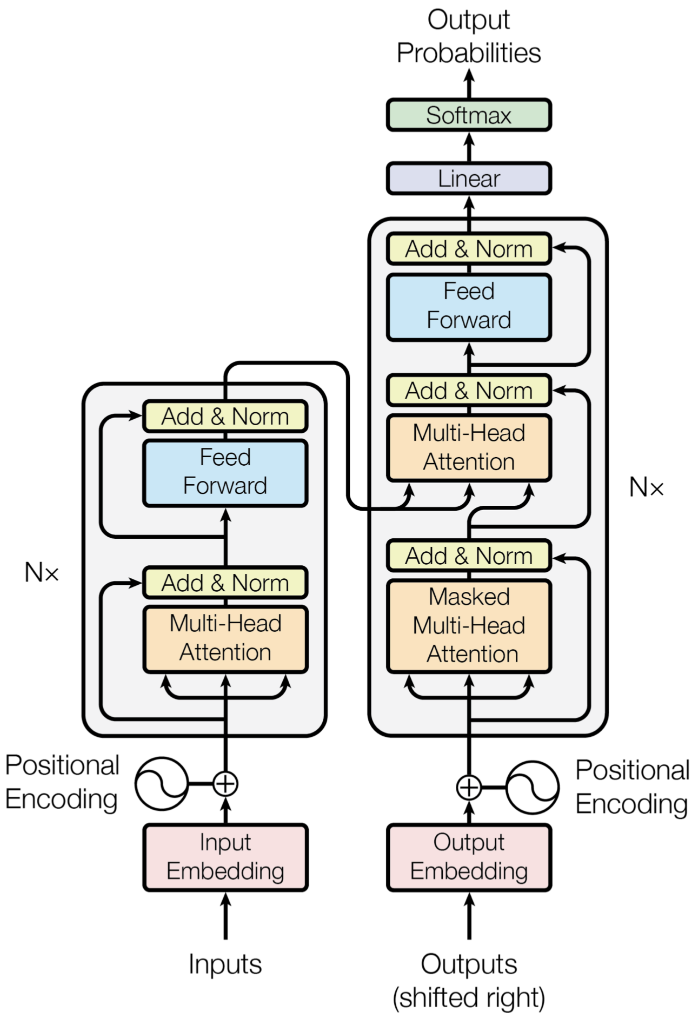


Today’s subject: Transformers (Decoder-only)



Recap of Transformer architecture

- The main components
  - Embedding
  - **Positional Encoding**
  - **Self-Attention**
  - Feed Forward
  - **Layer Normalization**
  - Residual Connections

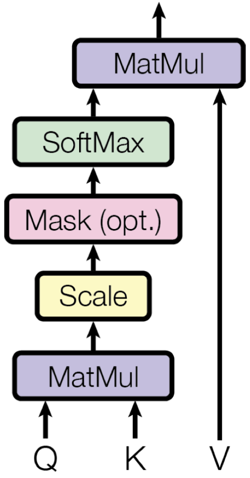


Recap of Attention mechanism

- Scaled Dot-Product attention
- where $\sqrt{d_k}$ is the dimension of the key vector k and query vector q

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Scaled Dot-Product Attention

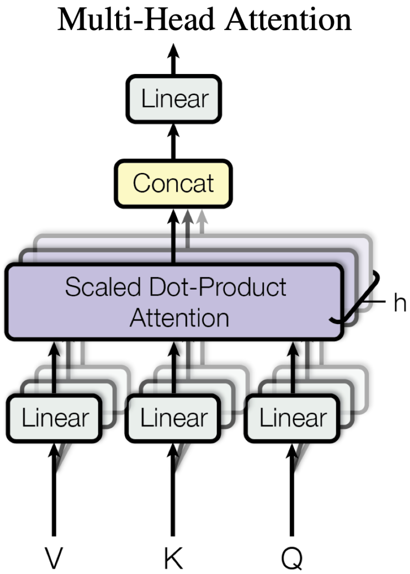


Recap of Attention mechanism

- Multi-head attention

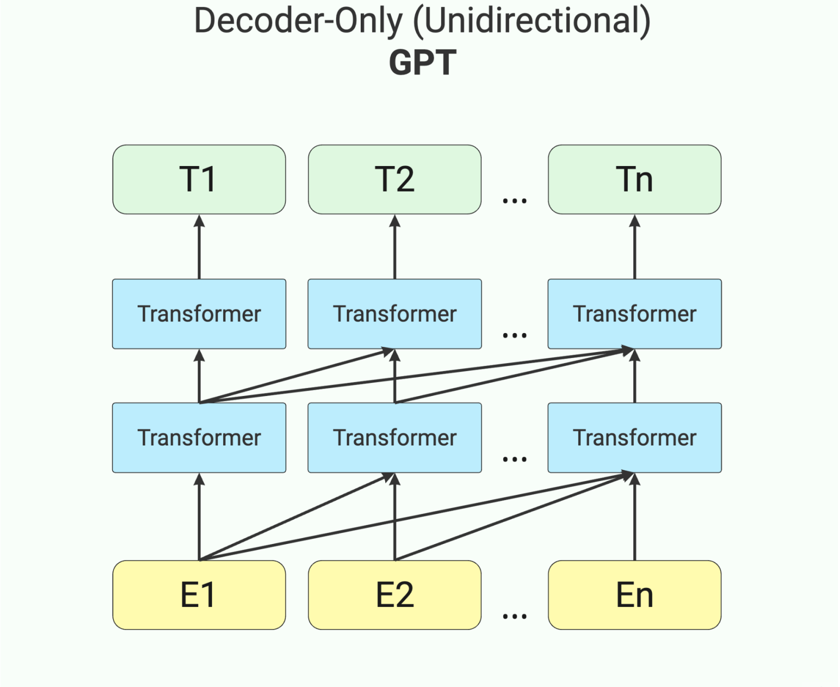
$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$



GPT: Decoder-Only (Autoregressive Architecture)

- **Decoder-only:** generates text by predicting the next word in a sequence based on prior context
- **Unidirectional:** processes text left-to-right, predicting one token at a time
- **Causal Language Modeling (CLM):** trains by predicting the next word in a sequence, avoiding future context
- **Self-attention:** focuses on relevant context in the input to determine the next output word
- **Pre-trained:** fine-tuned for specific tasks with minimal additional training
- **Text Generation:** excels at producing coherent and contextually relevant text based on prompts


Decoder-Only (Unidirectional)
GPT
T1
T2
...
Tn
Transformer
Transformer
...
Transformer
Transformer
Transformer
...
Transformer
E1
E2
...
En