Midterm review :

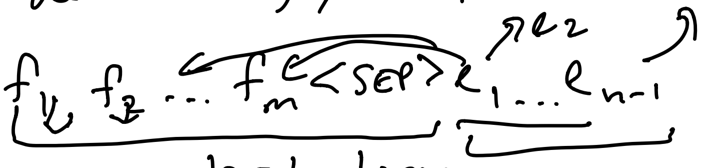
Important but non-exhaustive topic list :

- Language models
  - n-gram models
  - perplexity
  - simple neural LMs
    - fixed-window NLM
  - RNNs
    - not parallelizable at training time
  - Transformer LMs
    - self attn / cross-attn
      - query / key / value
      - masking
    - types of Transformers
      - decoder-only
      - encoder / decoder
        - encoder: compute representations of its input, which can be used to condition the decoder

$$p(e_n \mid e_{1\ldots n-1}, \boxed{f_{1\ldots m}})$$

↳ encoder

- cross attn
- residual connection
- prefix LM
  - decoder-only, modified mask



- training vs. test time

- Training language models
  - n gram: count / normalize
  - neural LMs: - gradient descent
    - backprop
    - cross-entropy loss
      used for next word prediction
  - batching
  - tokenization
    - words, characters, subwords, bytes
    - BPE
  - Adapting to downstream tasks
    - pretrain / finetune
      - BERT / T5

- prompt tuning

- Instruction tuning
    - FLAN
- RLHF

- Retrieval-augmented LMs
    - REALM

- Using LMs at test-time
    - decoding algorithm
        - greedy
        - beam search
        - sampling
            - ancestral / "pure" sampling
            - truncated sampling
                - nucleus, "top-p" sampling
    - prompting techniques
        - zero-shot / few-shot / instruction
        - "prompt engineering"
        - chain-of-thought
        - retrieval

- Evaluation of LMs
  - automatic eval metrics
    - perplexity,
    - BLEU for MT
    - ROUGE for summarization
    - BLEURT/COMET
  - human eval