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-att / cross-att
    - query/key/value
    - masking
- Types of Transformers
  - decoder-only
  - encoder/decoder
    - encoder: compute representation of its input, which can be used to condition the decoder

\[ P(e_n | e_1, ..., e_{n-1}, f_1, ..., f_m) \]
- cross att
- residual connection
- prefix LM
- decoder-only, modified mask
  \[ f_1, f_2, ..., f_m, <sep>, e_1, ..., e_n \]
- 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 / TS
- prompt tuning
- Instruction tuning
  - FLAN
- RLHF
- Retrieval-augmented LMs
  - REALM
- Using LMs at test time
  - decoding algorithms
    - greedy
    - beam search
  - sampling
    - ancestral / "pure" sampling
    - truncated sampling
      - nucleus, "top-k" 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