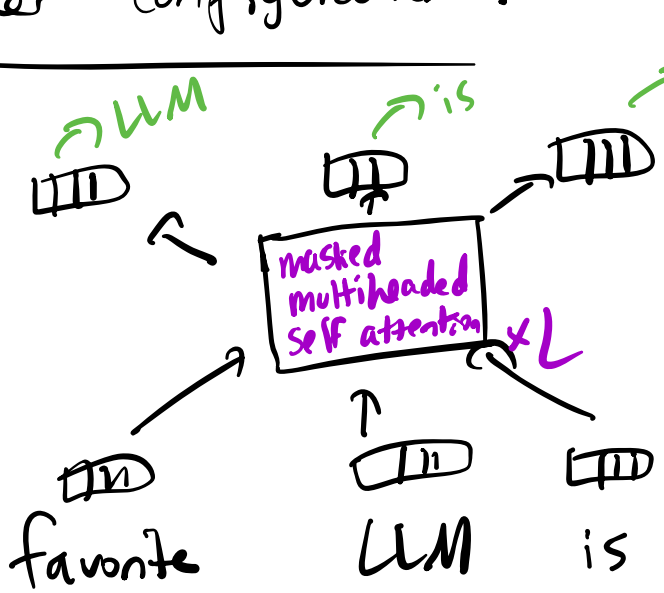
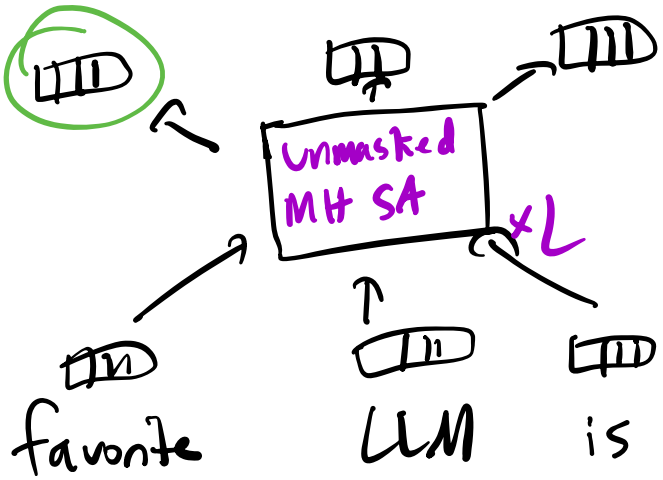
Transformer configurations:

Decoder:



⇒ useful for text generation

Encoder:



$(z_1^L, z_2^L, \ldots z_7)$

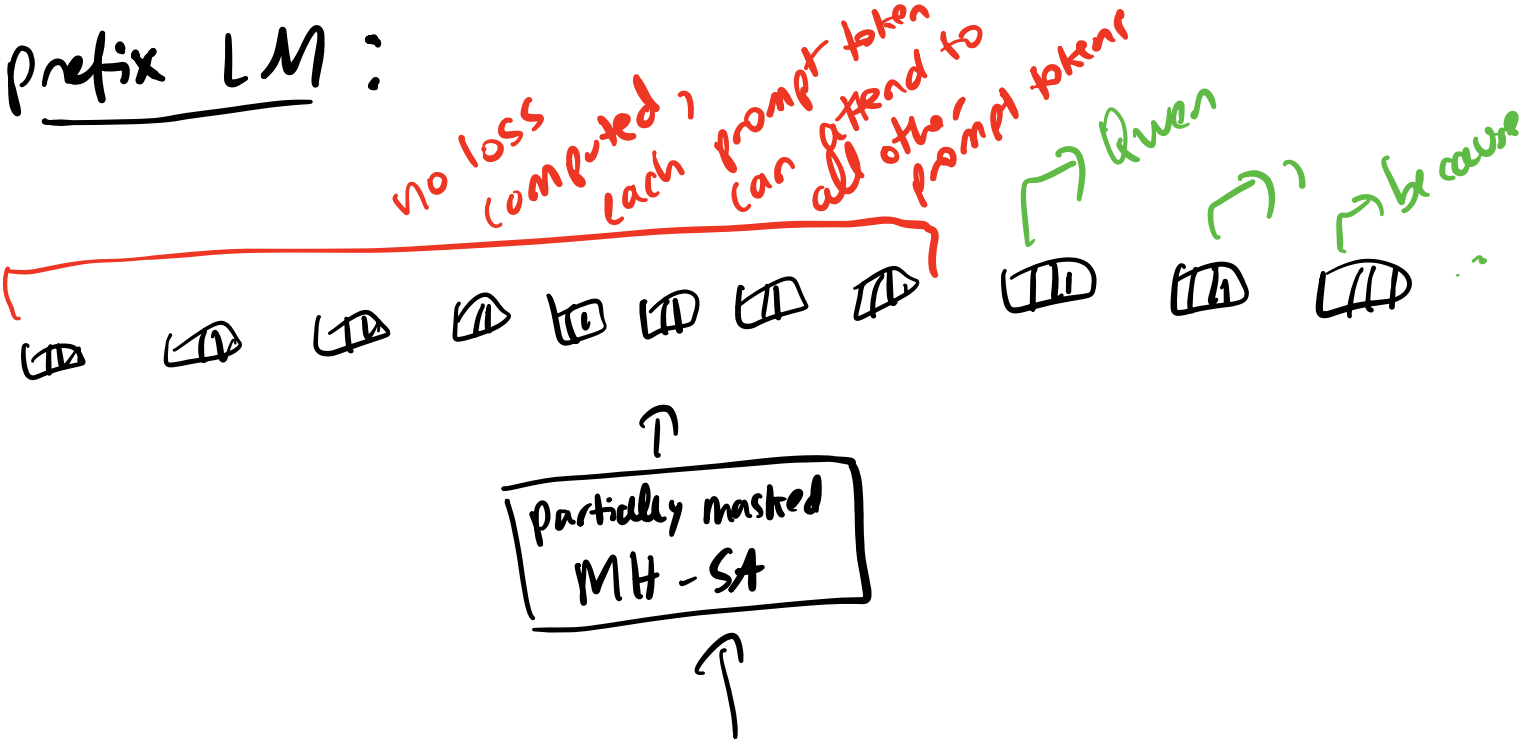
↳ useful for computing representations of an input sequence

↳ these reps can be used in other applications e.g. text classification

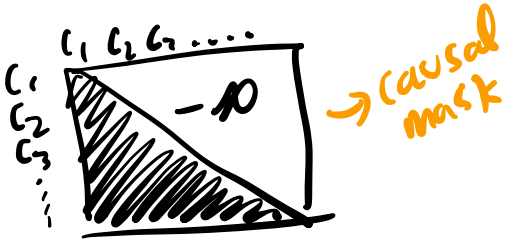
↳ cannot generate text from these models

↳ ex: BERT, RoBERTa, Electra, ... diffusion LMs

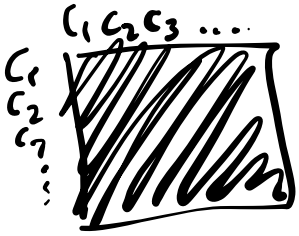
prefix LM:



decoder mask



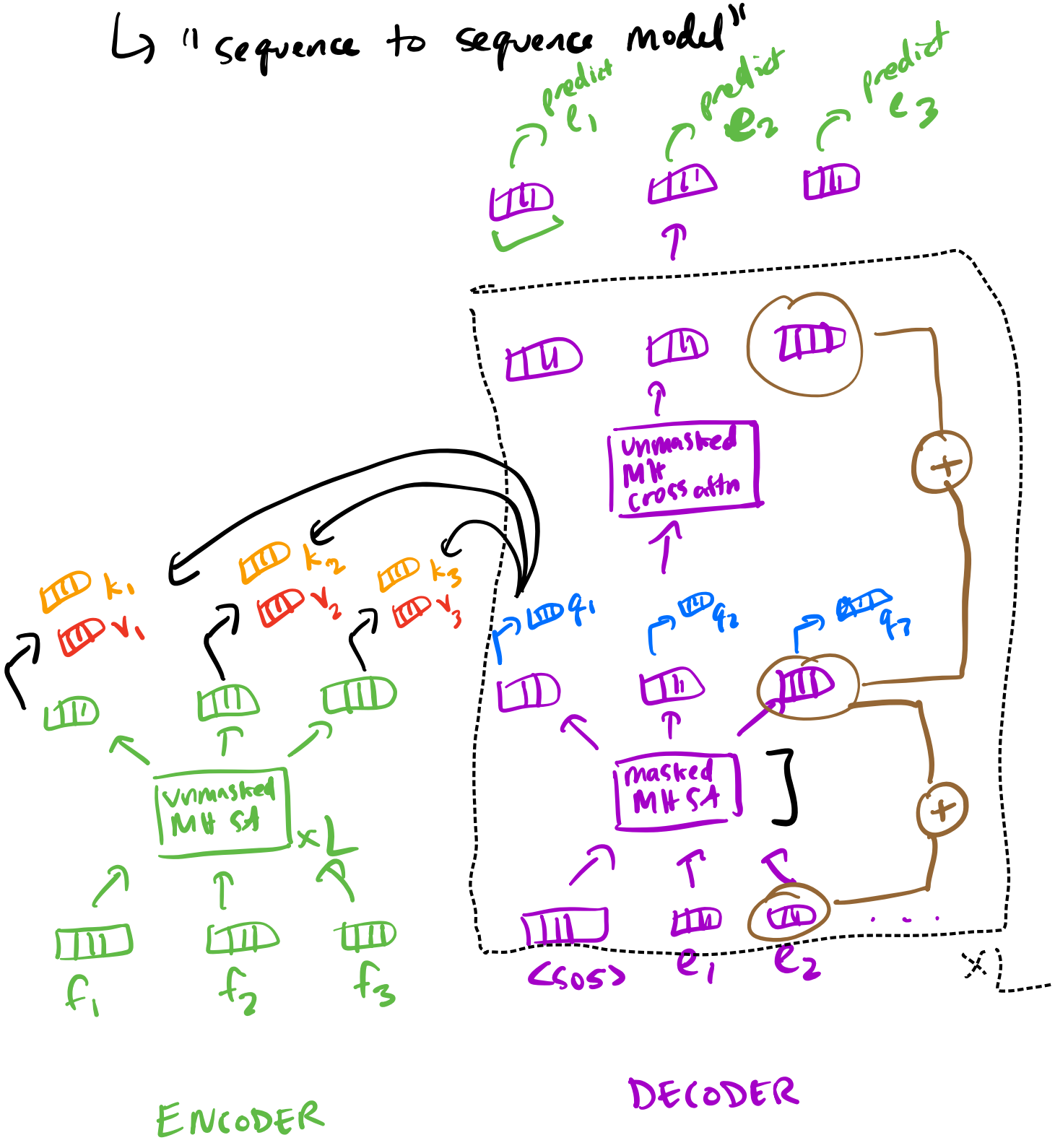
encoder mask



prefix LM

encoder/decoder model

↳ "sequence to sequence model"



↳ cross-attn always uses keys/values computed from the final layer z_i vectors of the encoder

Pretraining vs. post-training

- pretraining is conducted w/ as much text as possible
  - trillions of tokens
    - internet crawls (Common Crawl)
    - "high-quality" data
      - generally copyrighted
      - textbooks, novels
  - biggest model we can afford
  - goal: to obtain a model that "understands" many linguistic properties
    - grammar
    - world knowledge
      - "The President of France is ___"
    - emergent properties
      - in-context learning
      - chain-of-thought

↳ post-training

↳ goal:

1. make a pretrained model better follow instructions
2. align the model to human intents & values

↳ generally using less data than pretraining

↳ fine-tuning (SFT)
reinforcement learning