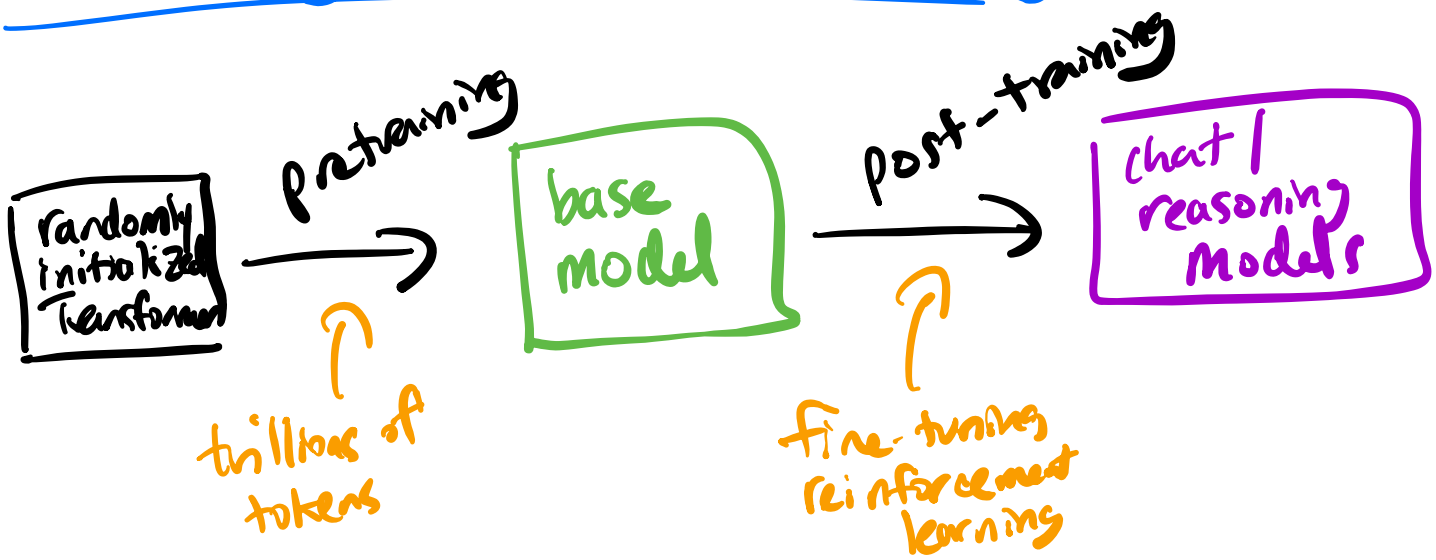
Pretraining vs. post-training



pretraining

- next-token prediction on a giant dataset
- train the biggest model you can
- goal: learn basic linguistic understanding, world knowledge, procedural knowledge,

post-training

- goal: make our model follow instructions
  - "capabilities"
- for fine-tuning, we use much less supervised "SFT"

data then for pretraining

↳ require input/output for SFT

↳ pretraining and SFT are very similar:

↳ same model, same loss fn

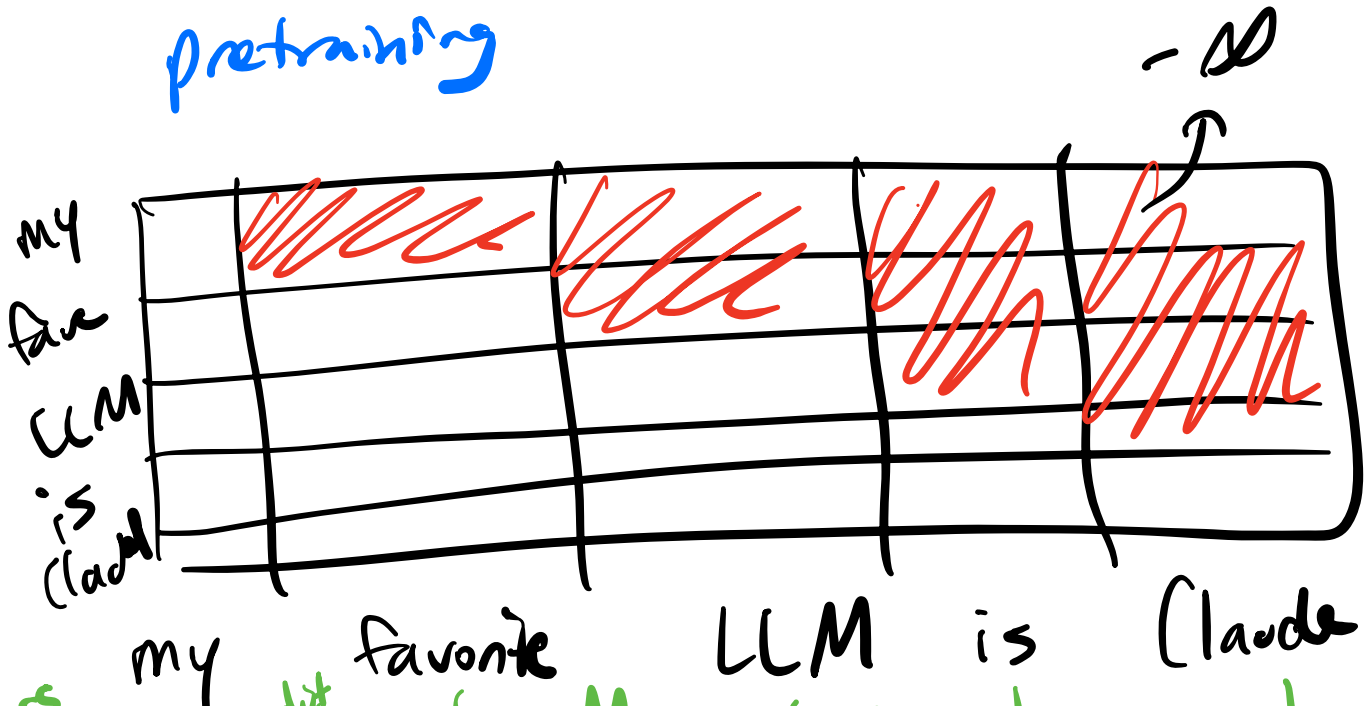
↳ different data, mask

---

difference in masking:

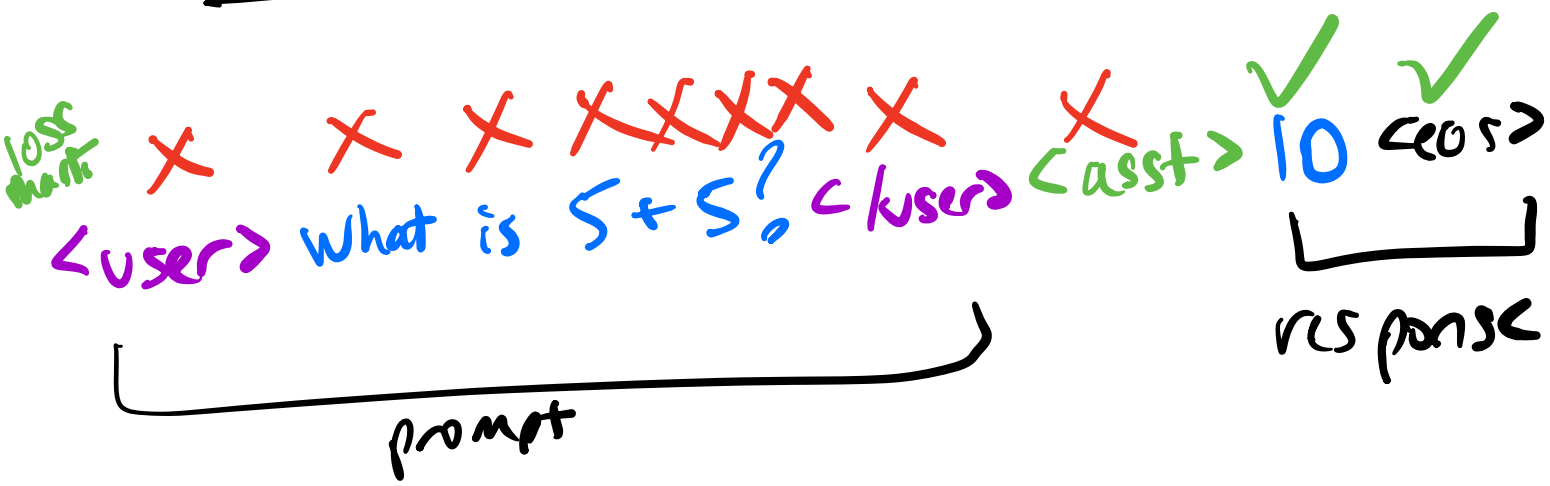
↳ attention mask

↳ loss mask



SFT:

Chat template: format for SFT data



↳ loss mask applied to train on only "assistant" tokens (aka the ground-truth response)

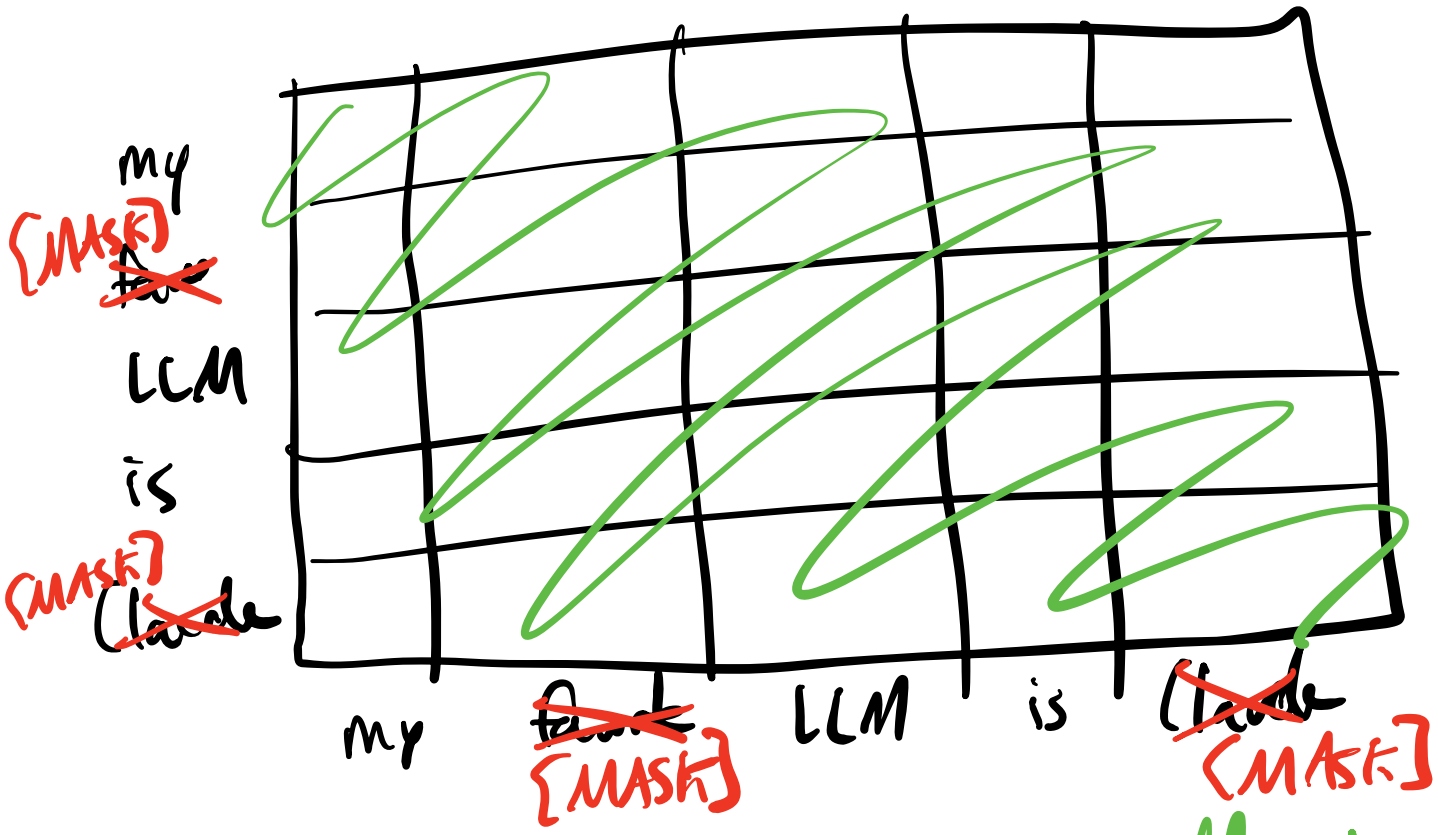
↳ everything else is unchanged from pretraining, the model still produces final layer hiddens z_i^L for all tokens

encoder Transformer

↳ BERT (2018)

↳ no attention mask

↳ specific loss mask based on input



↳ every token can see all other tokens (before and after)

↳ [MASK] is a special token

↳ replace k% of tokens in the input w/ [MASK]

↳ k generally 15-30%

↳ loss mask

loss mask ✗ ✓ ✗ ✗ ✓

my [MASK] LLM is [MASK]

↳ predict the identity of each masked token

↳ meant to be used when you want to embed a text

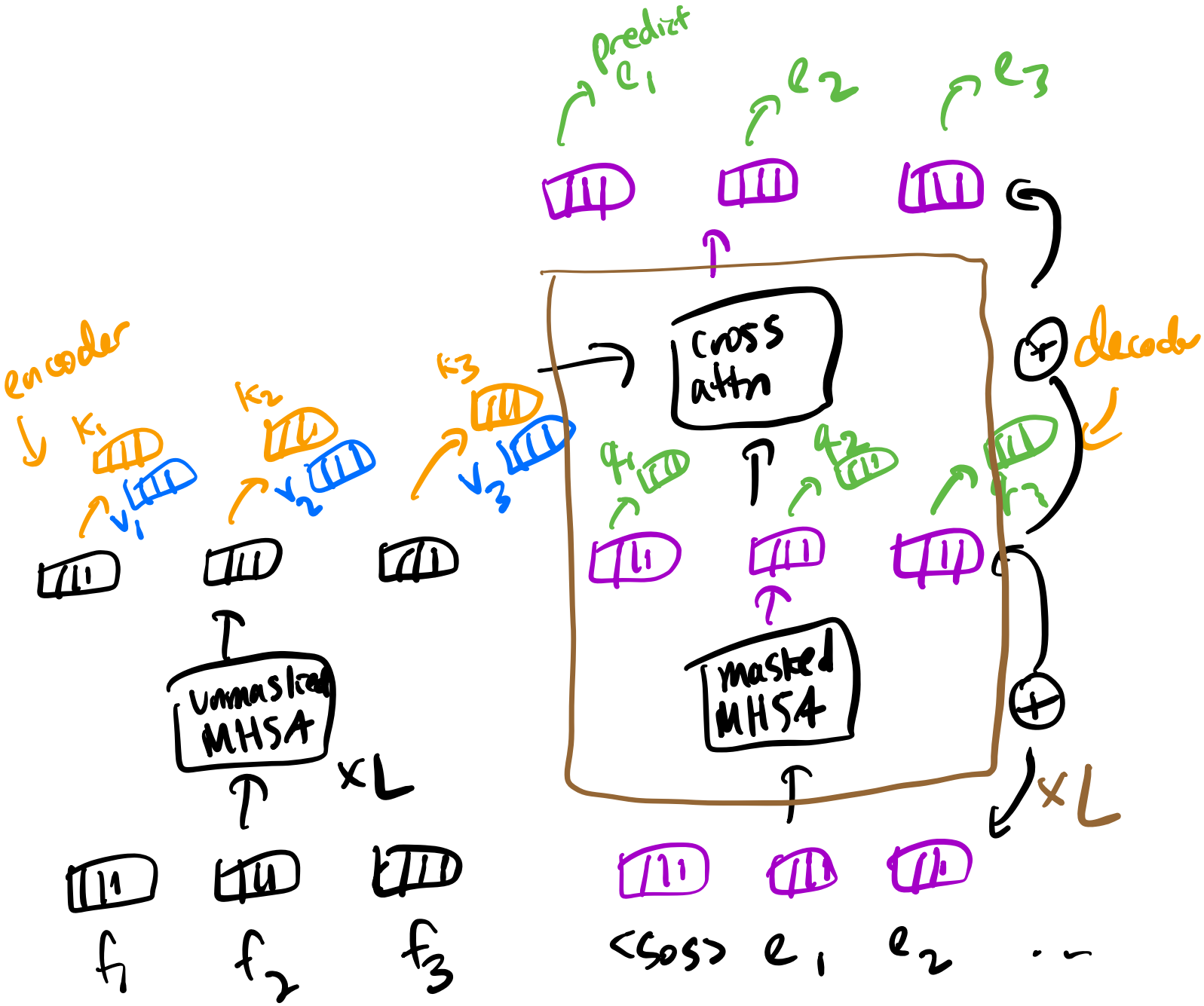
↳ clustering, information retrieval, simple classification

↳ more recently: diffusion language models, JeV

↳ encoder - decoder models

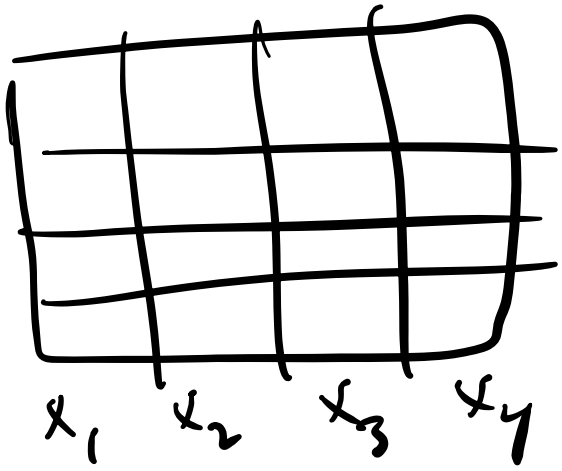
↳ machine translation

↳ translate French sentence $f_1 f_2 f_3$ into Eng $e_1 e_2 e_3$

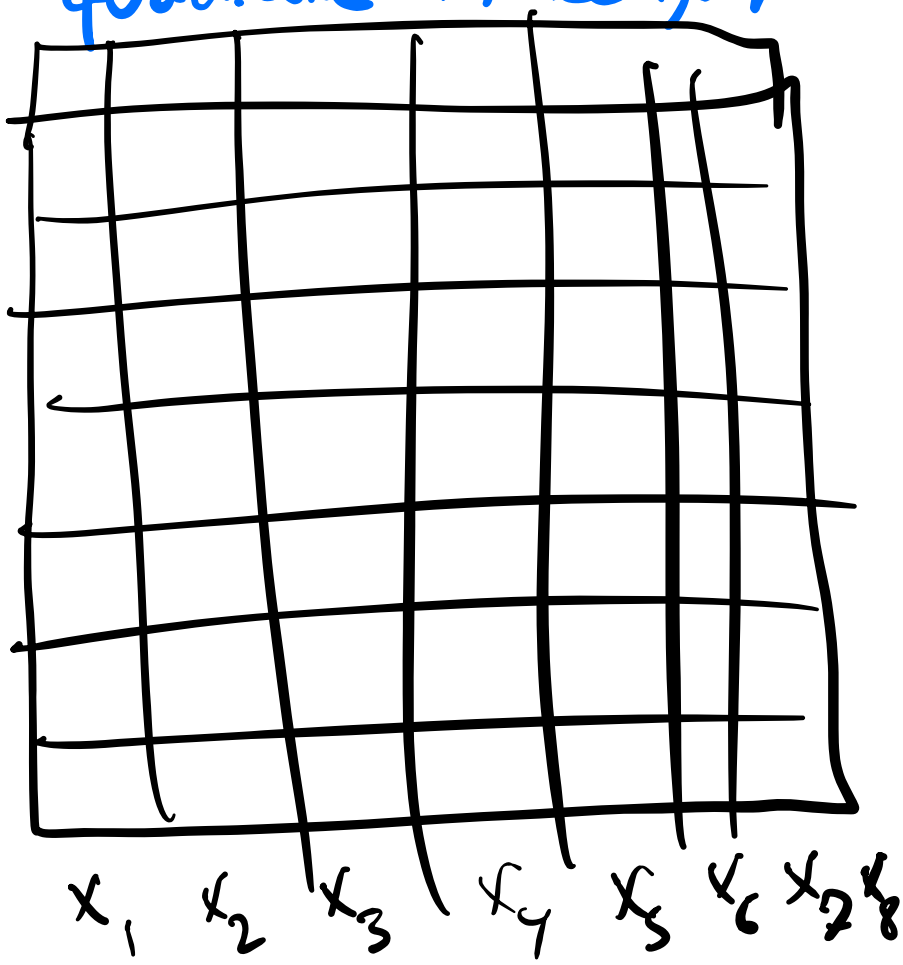


↳ loss only computed on decoder outputs

Self-attention is quadratic in length



length 4: 16 dot products



length 8: 64 dot products

↳ quadratic complexity motivates more efficient attn mechanisms