Final projects

- language-related
  - code / multimodal okay
    - language + { vision, audio, interaction

- high-level topic ideas
  1. "improve LM on task X"
     - open-weight LMs
       - Llama, Qwen, 1-7B param
     - collect a dataset
     - run baselines
     - train LM on task X
       - finetuning { LoRA, QLoRA
       - RL
       - prompt optimization
     - measure improvement
     - detailed error analysis
       - systematic error types

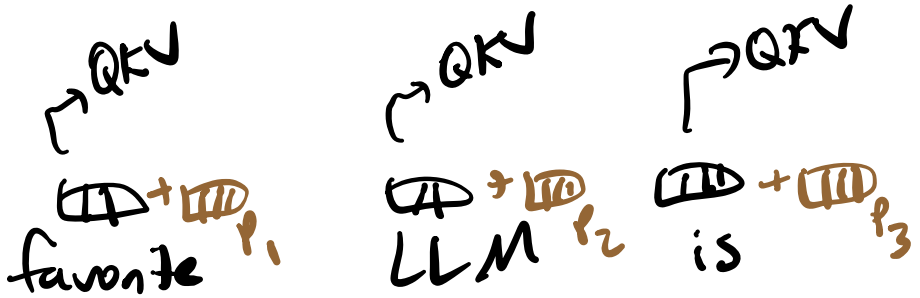
- "tool-using LLM agents"
  - build a system that interacts w/ the external environment
  - generally using closed models via API
    - try open models
  - evaluation is challenging
- "interpreting LM's behavior on some task"
  - identify layers / neurons / circuits
  - interventions
    - model unlearning, steering
- "implement new training method for LMs"
- "reproduce a recent paper"
- efficient inference
  - increase decoding speed but preserve quality
  - KV cache optimization
- LLM safety / security / jailbreaking

Self-attention:

↳ training time: parallelize hidden state computation

↳ test-time: sequential

↳ we can only parallelize computations when we know the full sequence ahead of time



Self-attn requires injecting position info into the input

KV cache

↳ storing prev. computed keys/values so we don't have to recompute them at every step

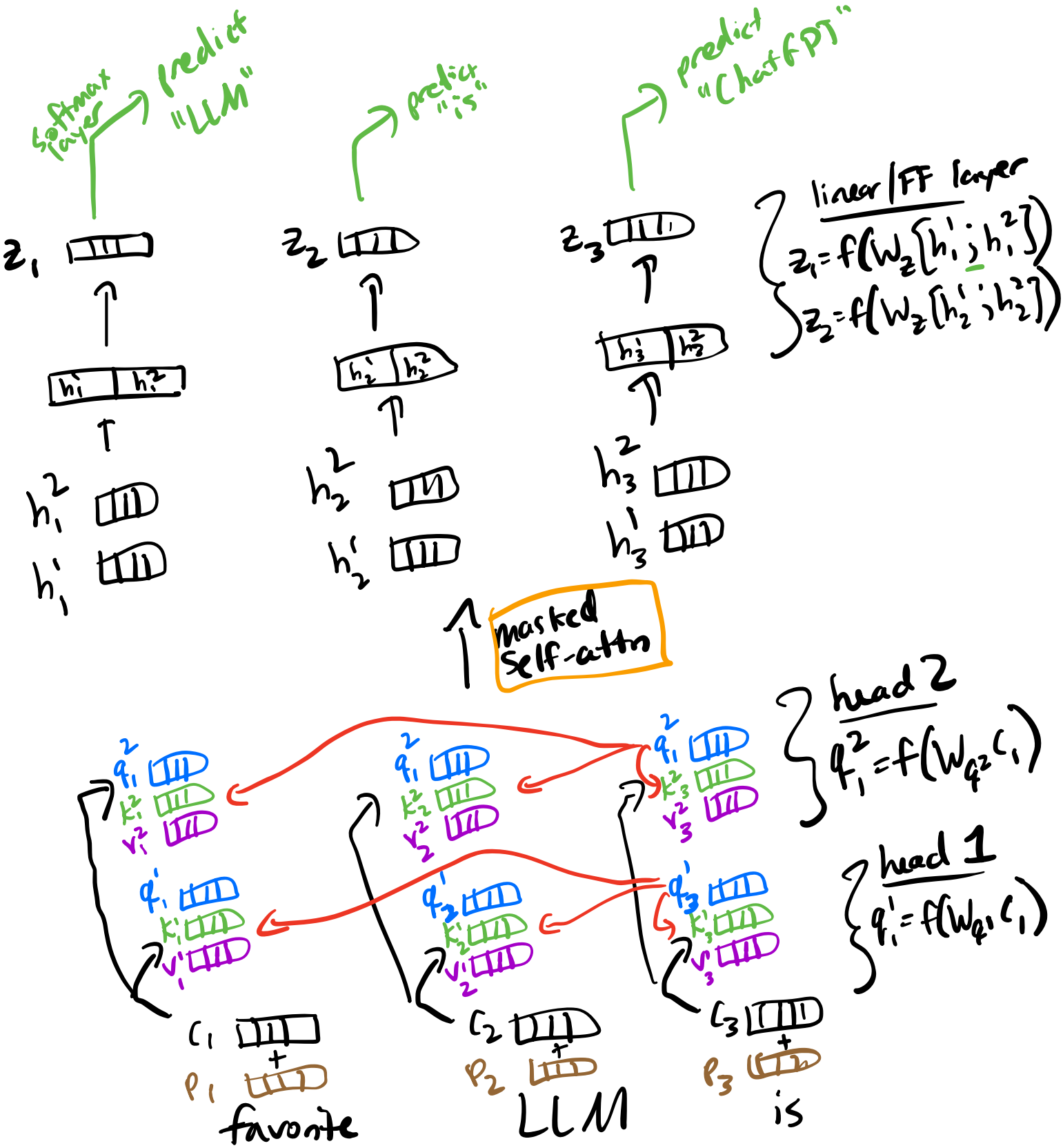
↳ test-time

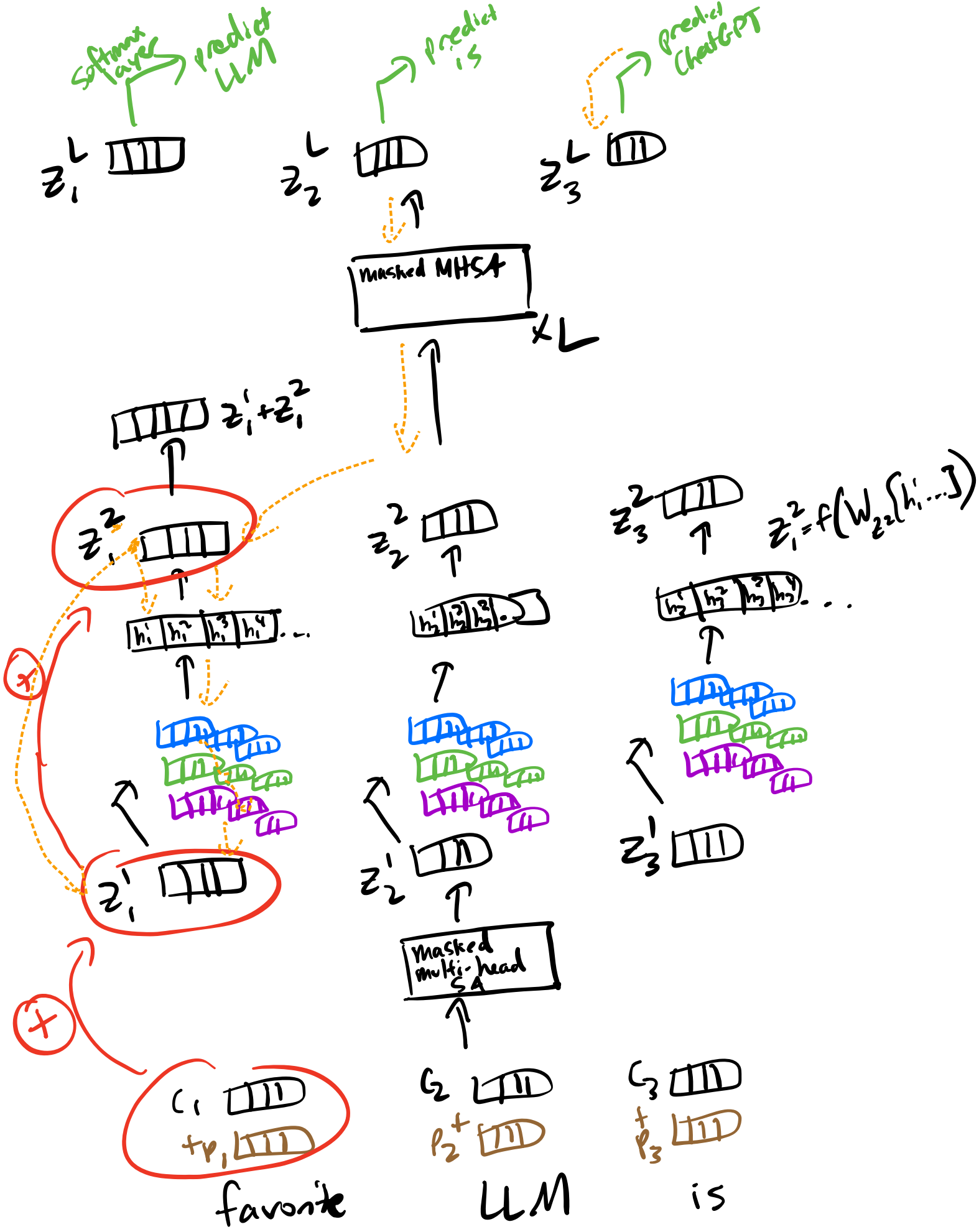
Transformer

- neural LM built on multi-head self-attn
- deep network with many stacked MHSA layers ("blocks")
- Vaswani et al. 2017

intuition:

- let's have multiple sets of Q, K, V for each token
  - then, each set (or "head") can learn to focus on a specific linguistic property
    - syntax (e.g. all verbs in prefix)
    - activate on certain words/phrases
    - entities/dates
    - position


softmax layer
predict "LLM"
predict "is"
predict "ChatGPT"
linear/FF layer
$z_1 = f(W_z[h_1^1; h_1^2])$
$z_2 = f(W_z[h_2^1; h_2^2])$
masked self-attn
head 2
$q_1^2 = f(W_{q^2} c_1)$
head 1
$q_1^1 = f(W_{q^1} c_1)$
favorite
LLM
is


softmax layer
predict LLM
predict is
predict ChatGPT
masked MHSA
xL
favorite
LLM
is