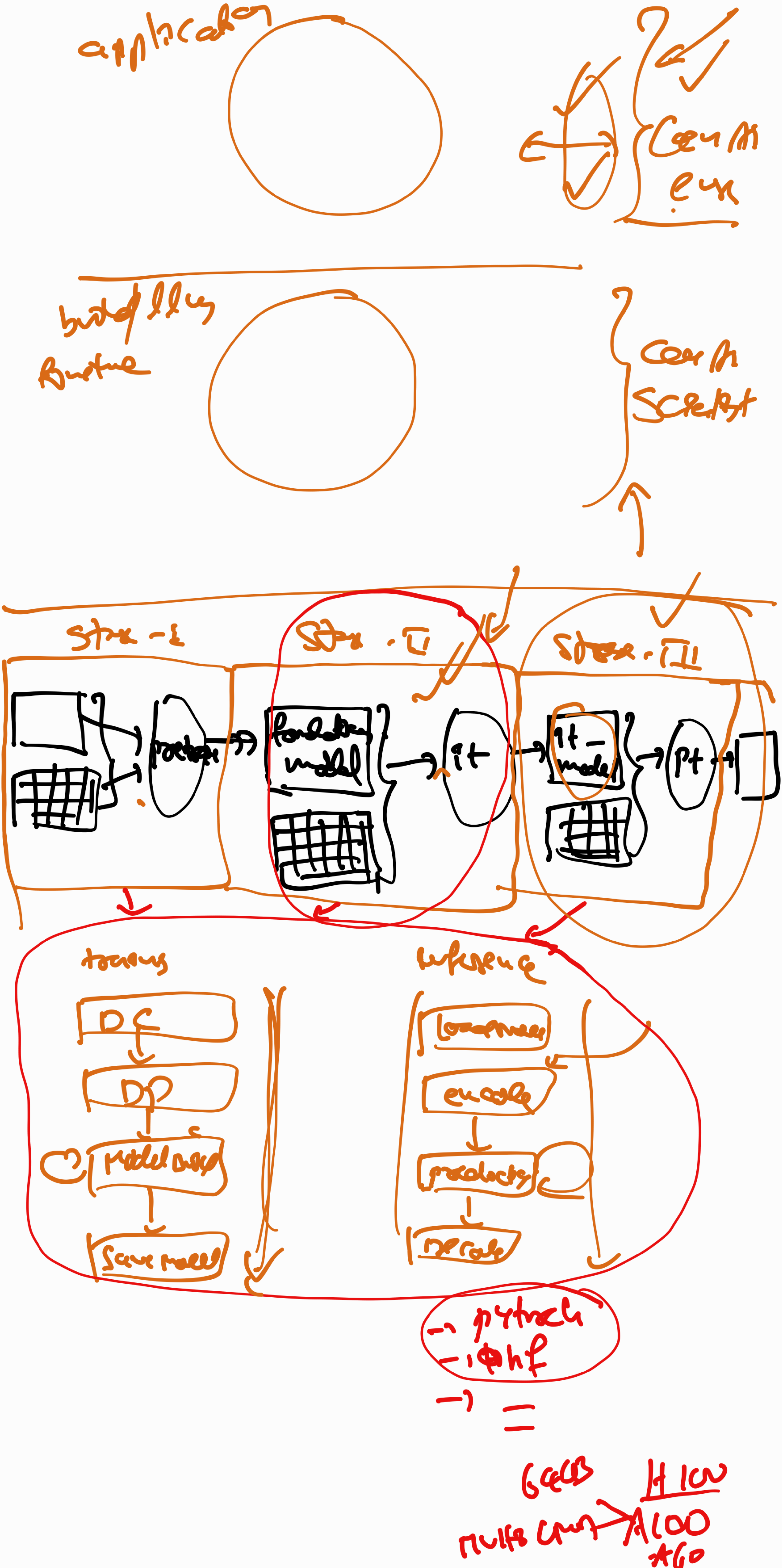
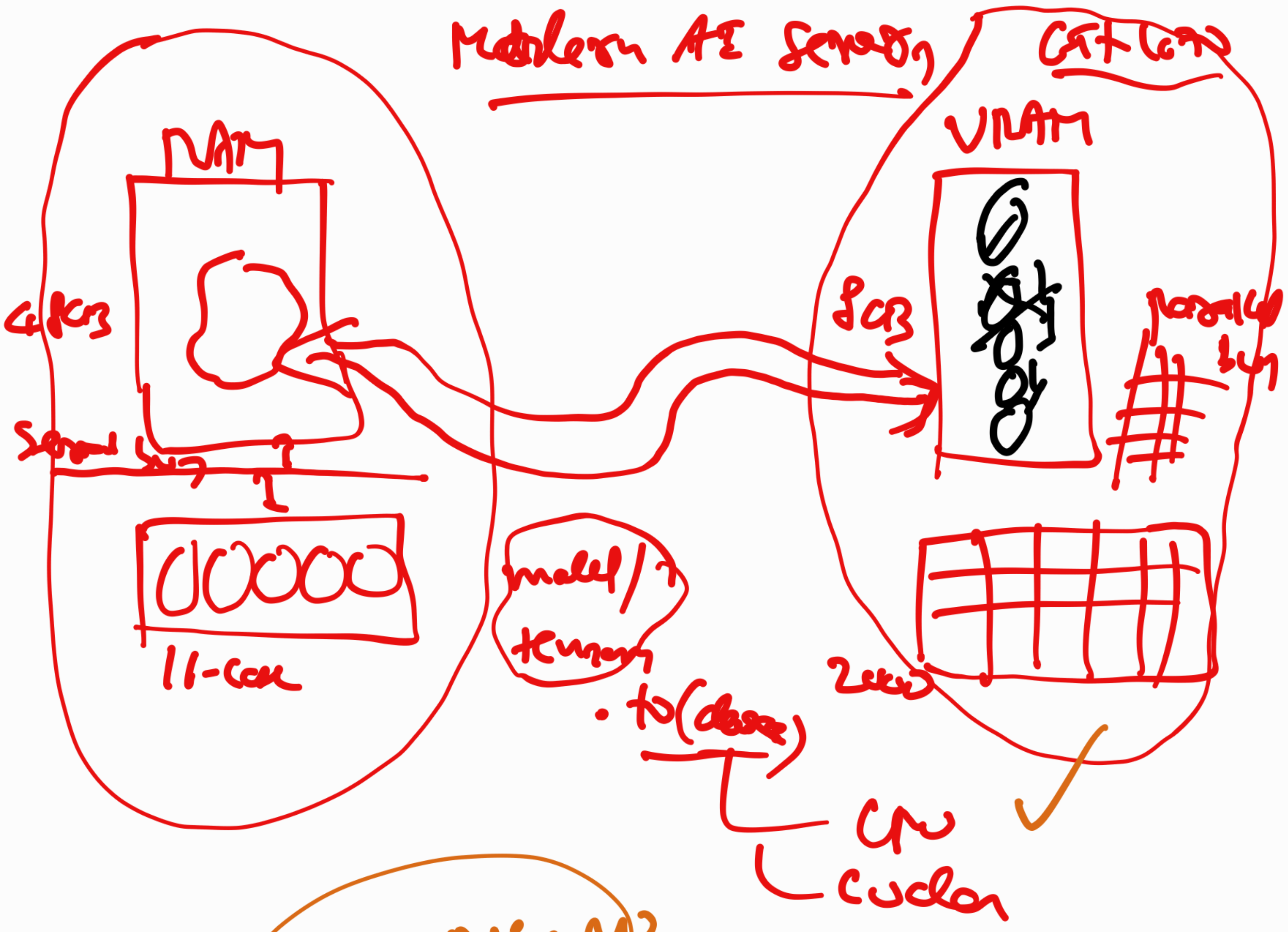
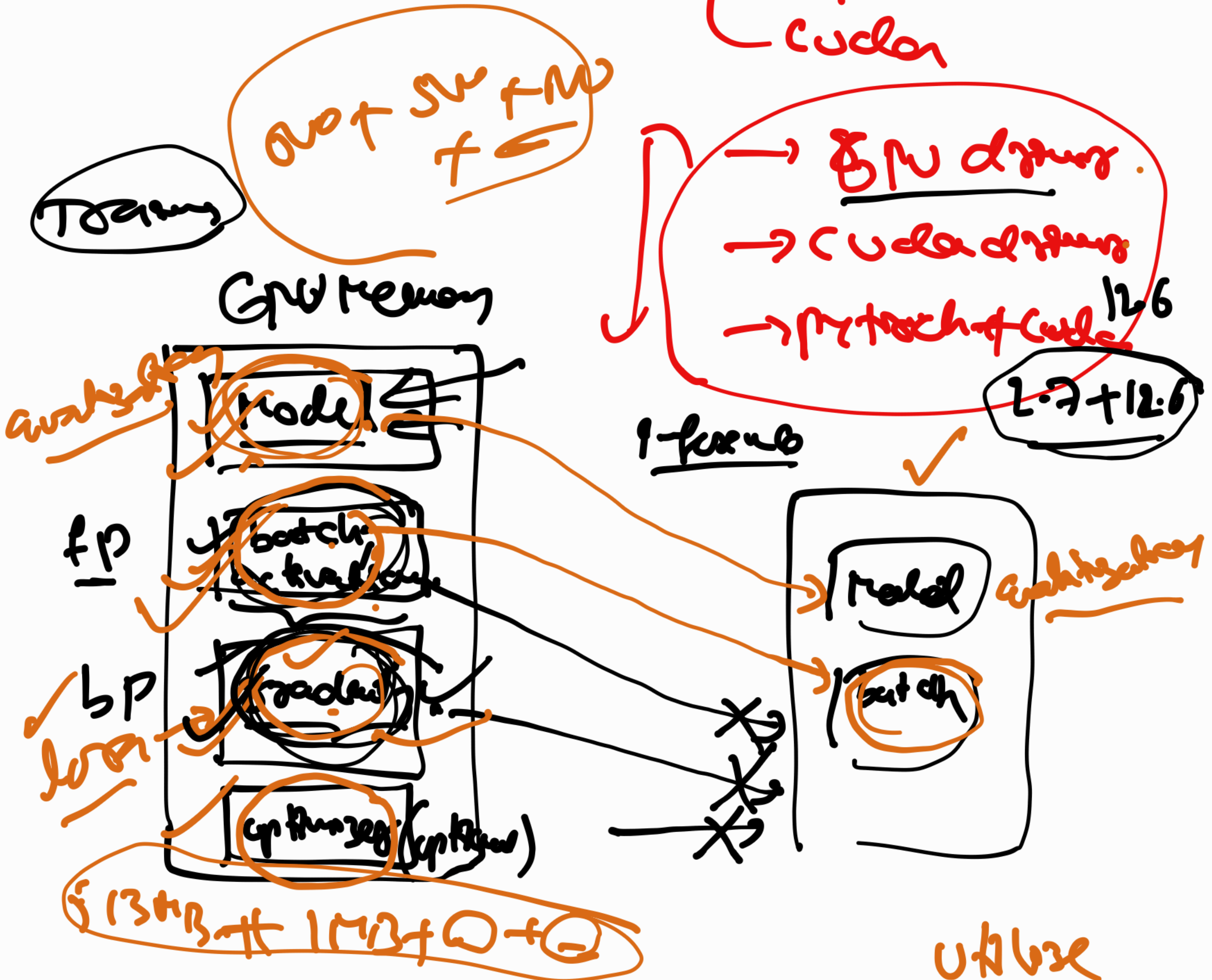


application
Stage - I
Stage - II
Stage - III
foundation model
training
inference
DP
save model
encode
pytorch
hf
H100
A100





efficiency techniques to utilize / optimize GPU memory effectively

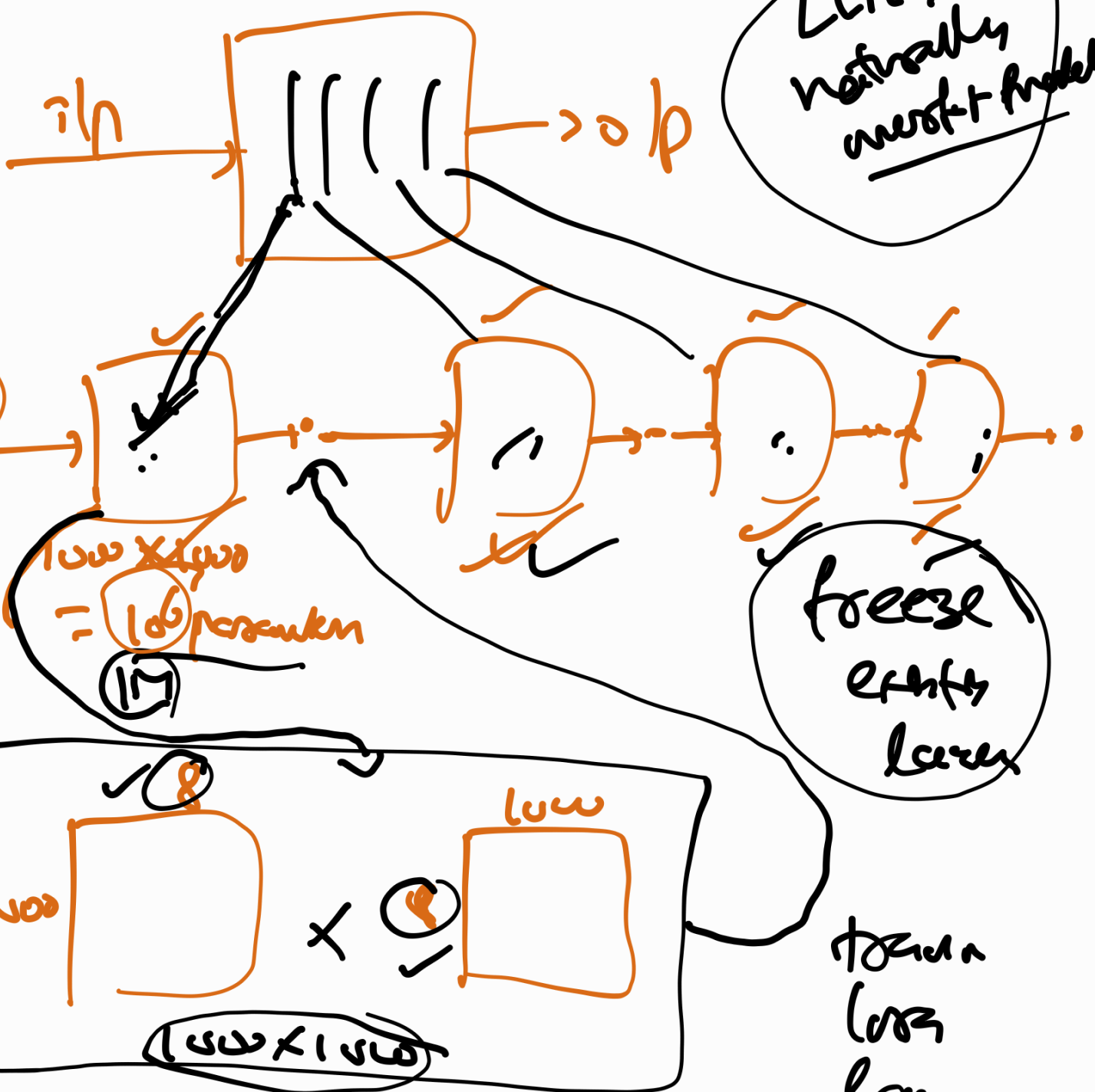
- LoRA
- quantization
- Liger kernel
- flash attention
- caching

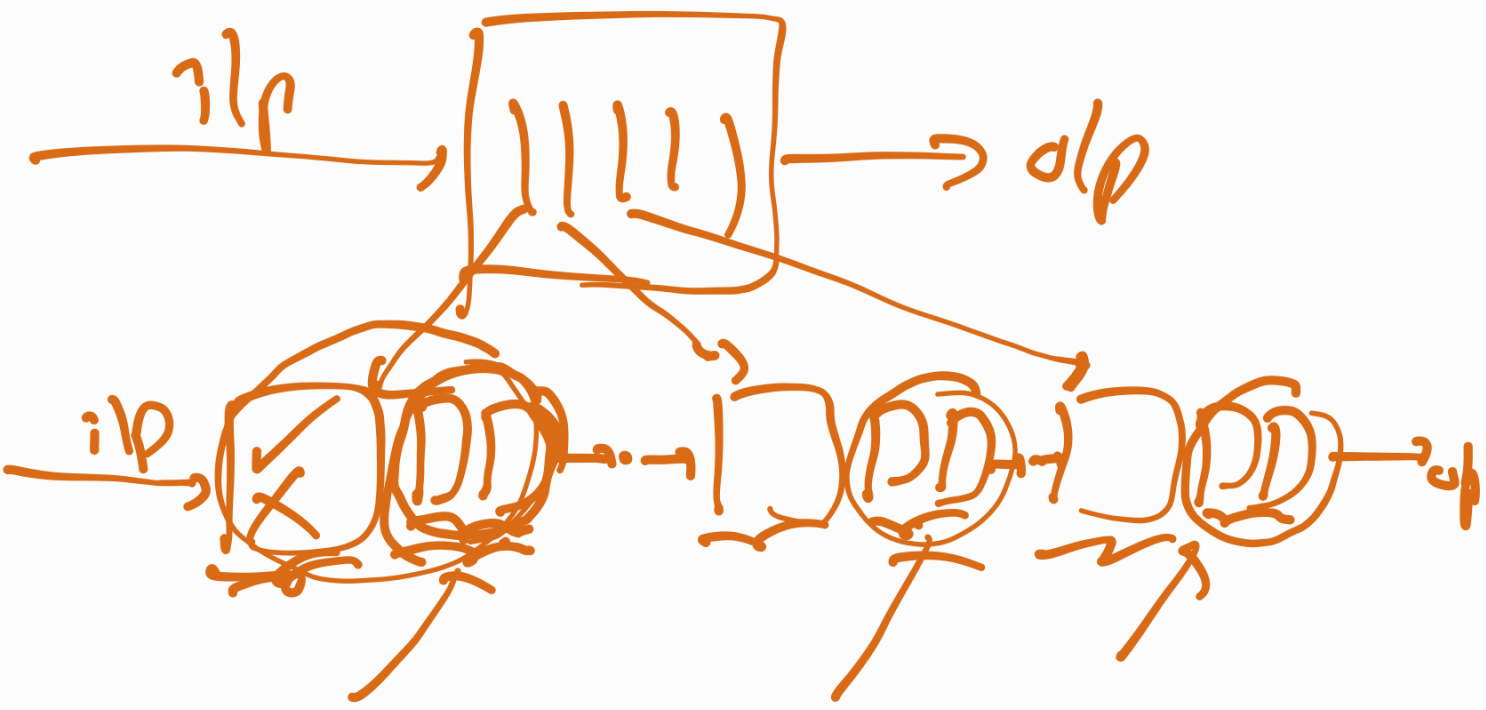
2023–24

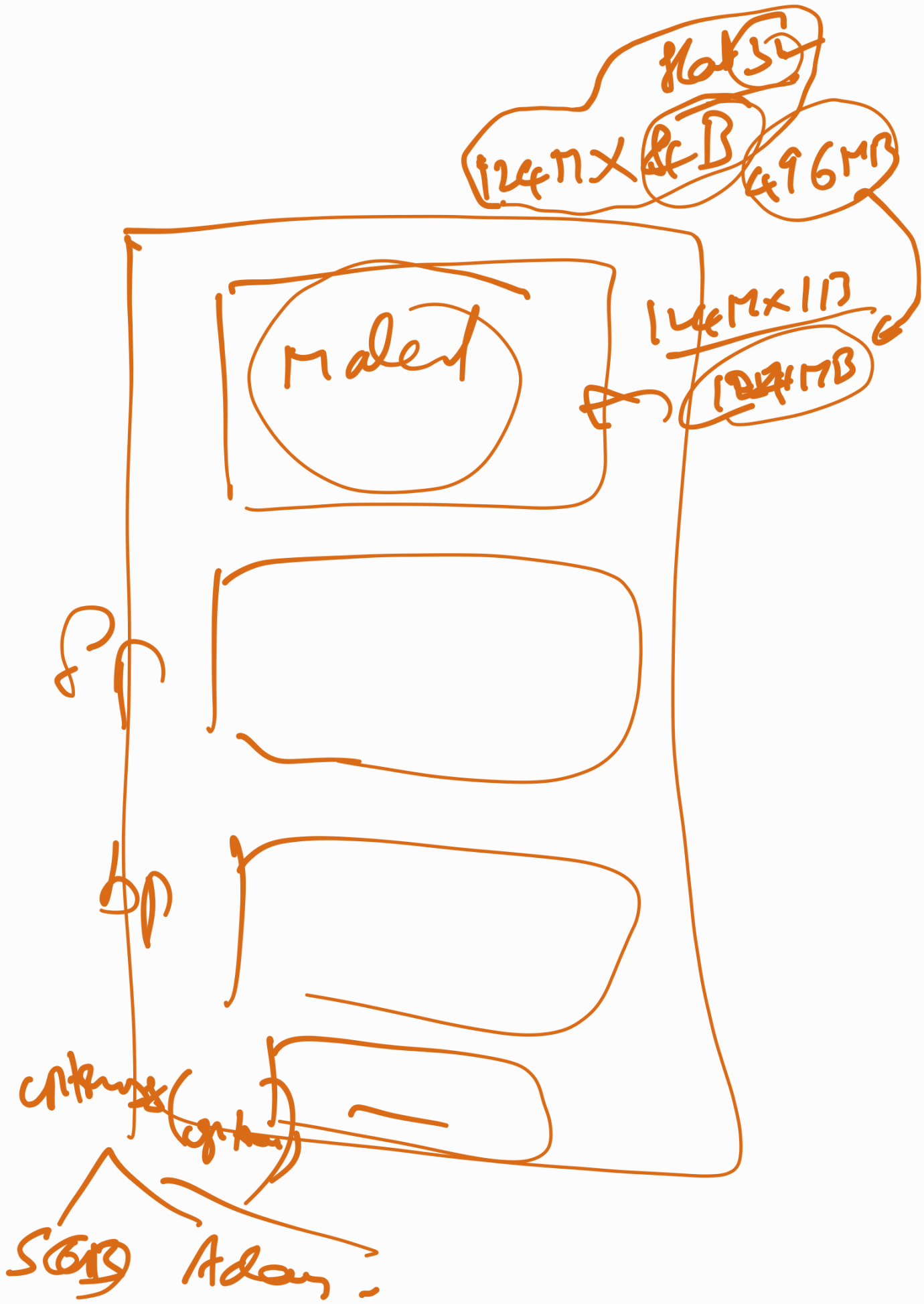
MoE

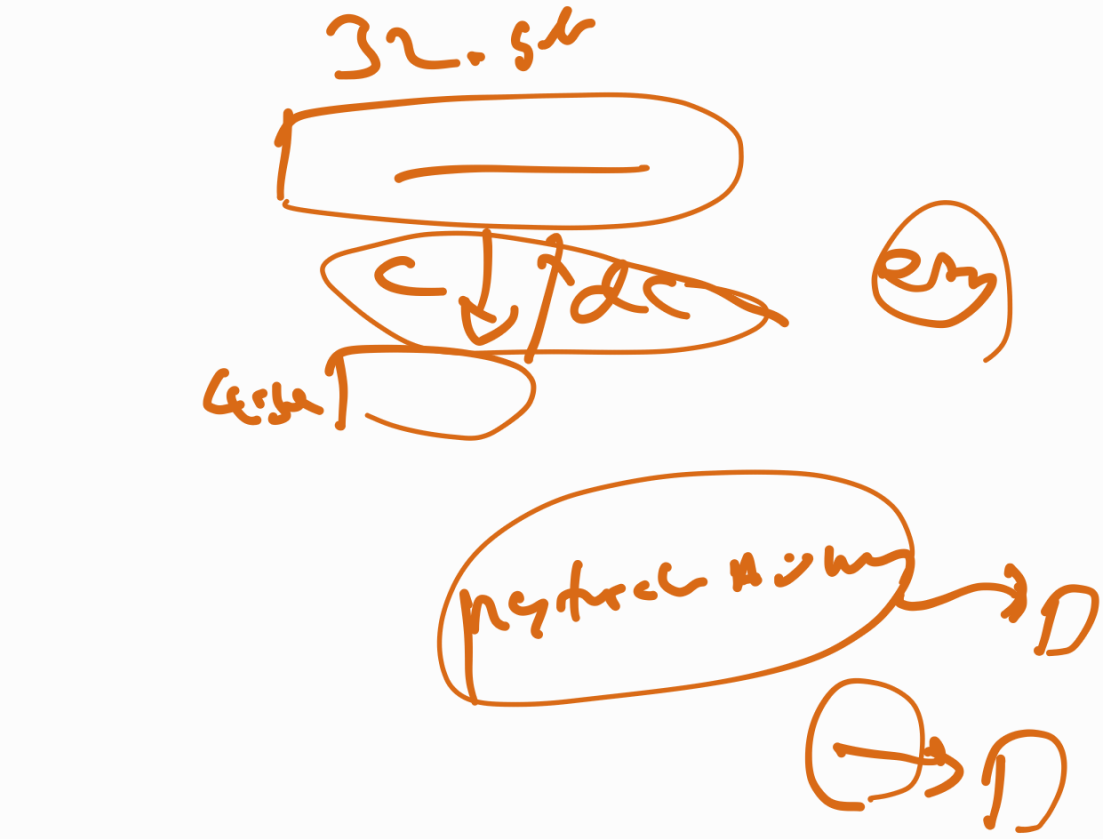
distillation

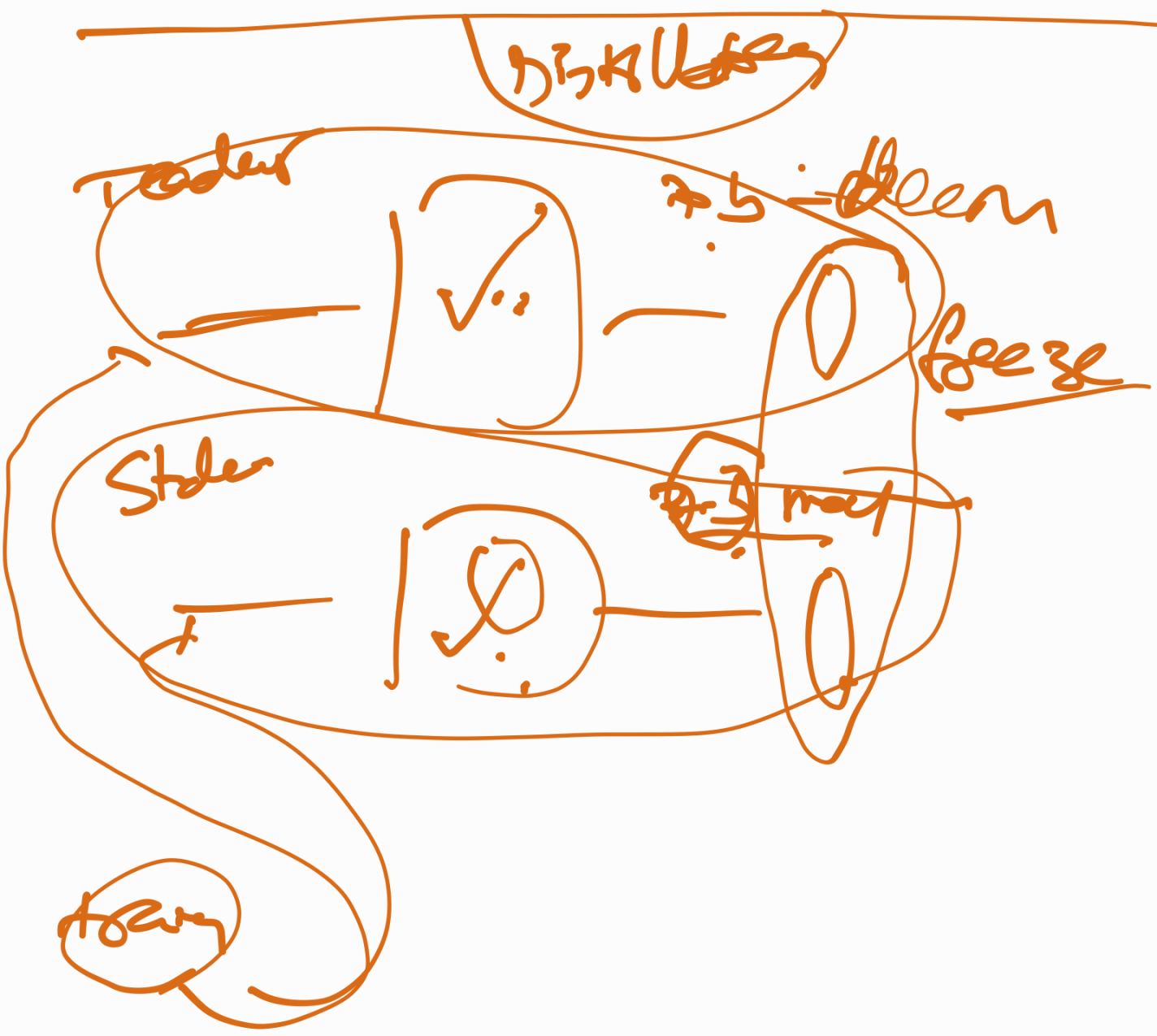
Low Rank Adaptor (LoRA)

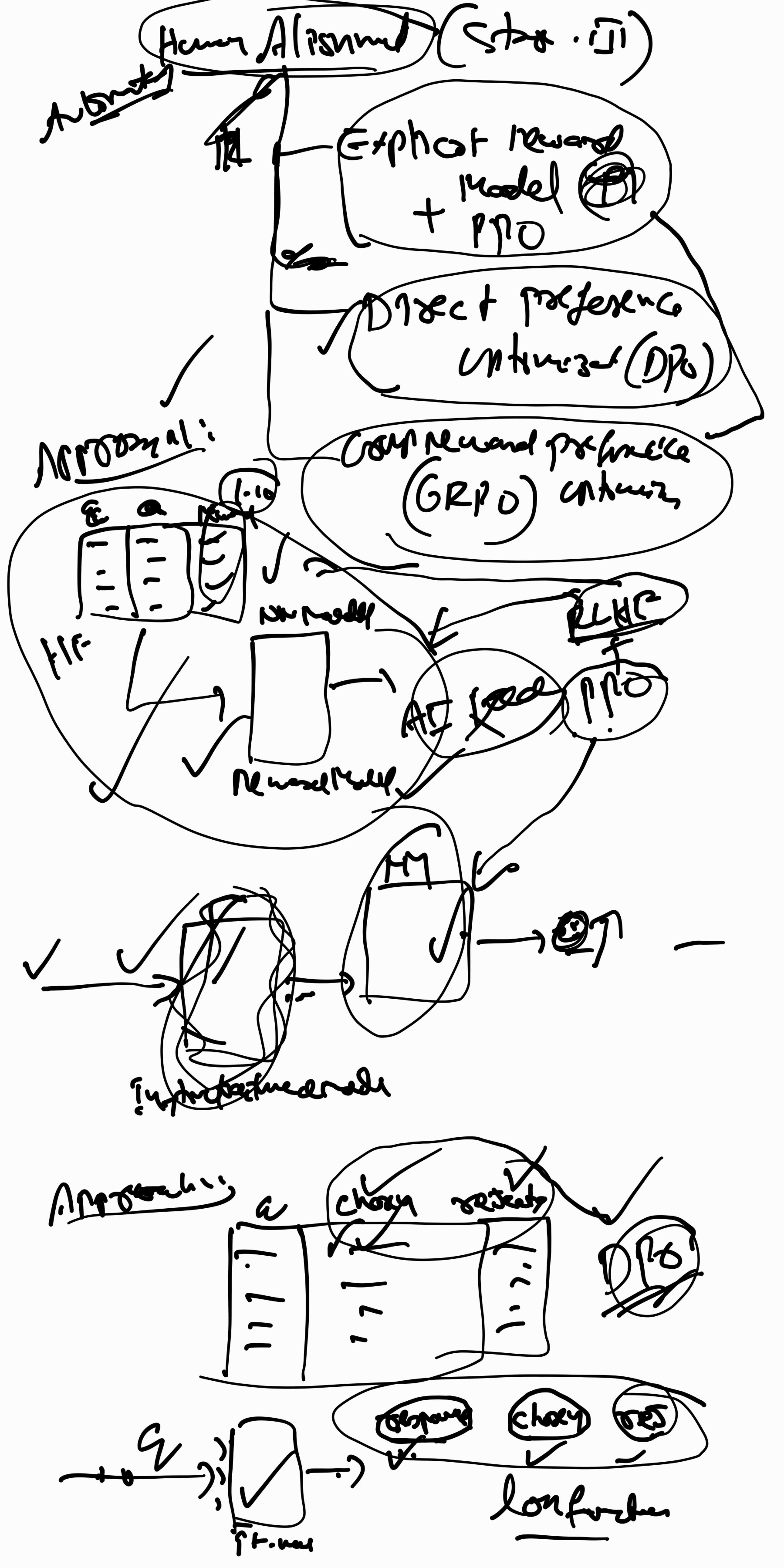





model
SGD
Adam






Human Alignment
Automated
RL
Explicit reward model + PPO
Direct preference optimization (DPO)
(GRPO)
Approach:
HF
NN model
Reward model
AI feed
RLHF
PPO
LLM
Approach:
Q
chosen
rejected
DPO
chosen
loss function

Approach 3: