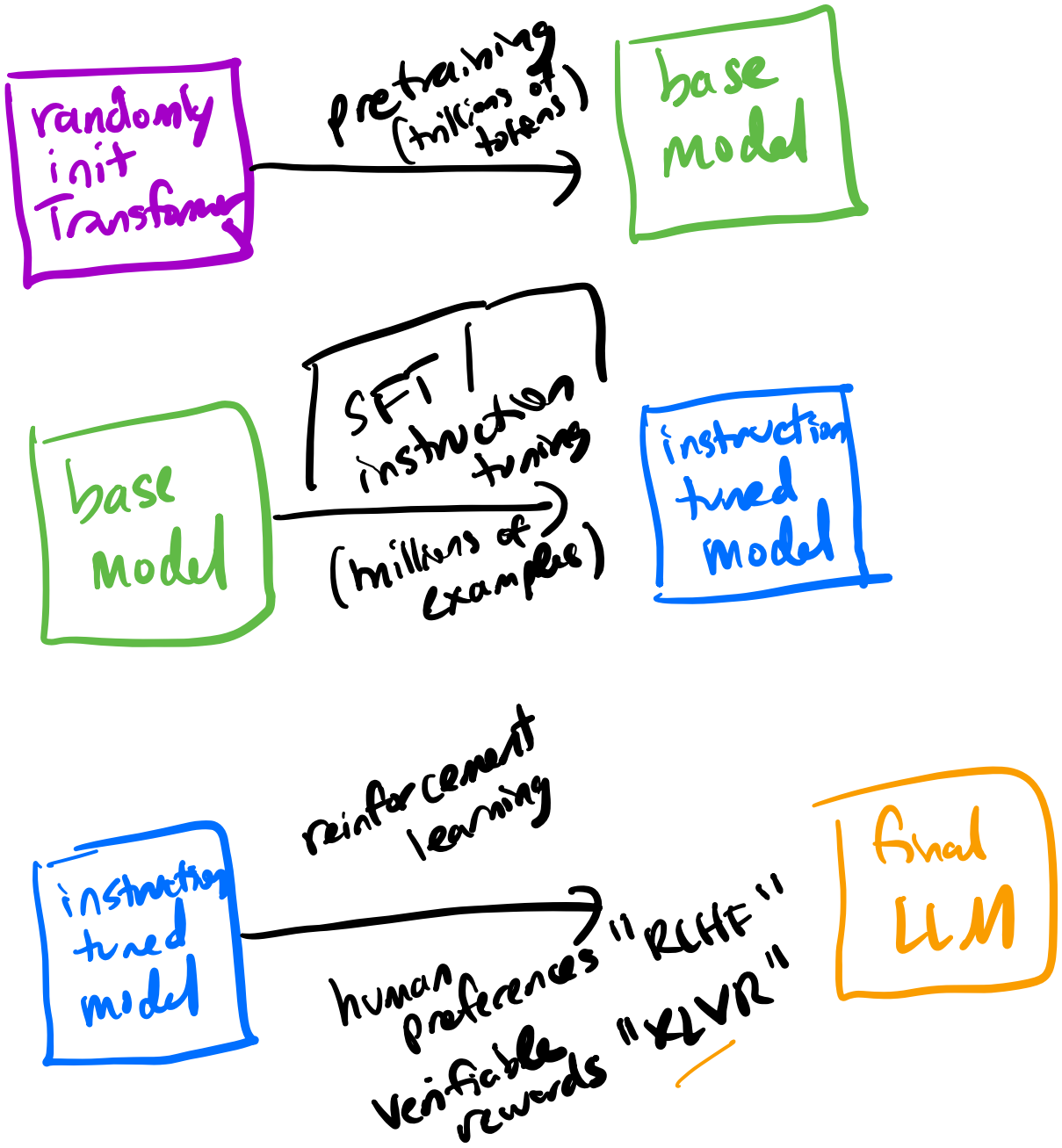
how can we get language models to follow our instructions?

1. in-context learning
   - ↳ add demonstrations (or "shots") to the prompt
   - \+ no extra training
   - − generally pretty bad
   - ↳ emergent property
   - ↳ 0 shot, 1 shot, few-shot, many shot

2. post-training
   - ↳ adapt the pretrained model by further training
   - ↳ common algos: SFT, RL, distillation


randomly init Transformer
pretraining (trillions of tokens)
base model
base model
SFT | instruction tuning
(millions of examples)
instruction tuned model
instruction tuned model
reinforcement learning
human preferences "RLHF"
verifiable rewards "RLVR"
final LLM

SFT:

- input/output pairs for one or more tasks
- minimize NLL on outputs (loss masking)
- just like pretraining, different data + loss masking
- SFT on one task
  - risks "catastrophic forgetting"
    - model gets worse at everything except the task it is SFT'd on
  - fix: mix in batches of pretraining

instruction tuning:

↳ a more specific case of SFT, where the data contains a huge diversity of tasks

add these numbers: 5, 2, 7, A: 14
Solve the attached HW: A: see ....
make this email clearer... ----

↳ if I train on enough unique tasks, will the resulting model generalize to tasks it's never seen?

↳ FLAN (Google, 2021)

↳ bigger models trained on more data exhibit stronger generalization to unseen tasks