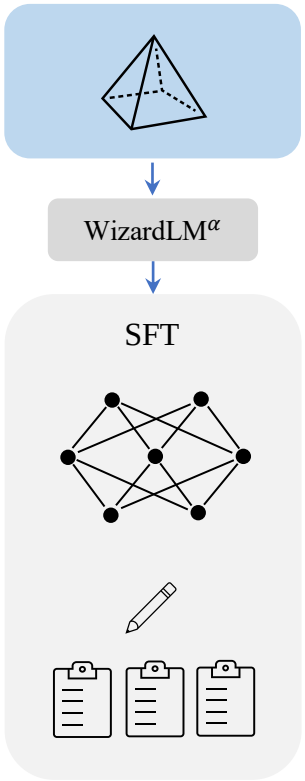
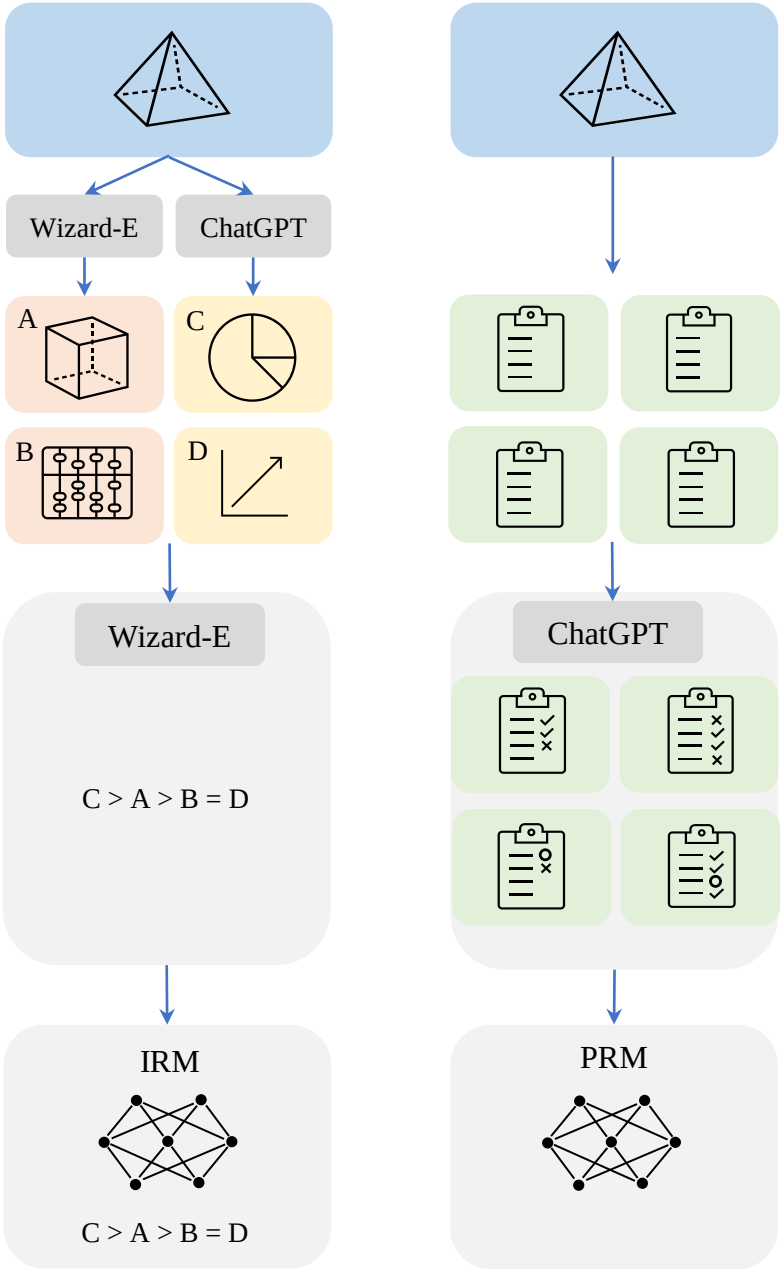
Step 1:
Supervised fine-tuning.



Step 2:
Training Instruction Reward Model (IRM), and Process-supervised Reward Model (PRM).



Step 3:
Active Evol-Instruct, and PPO training.