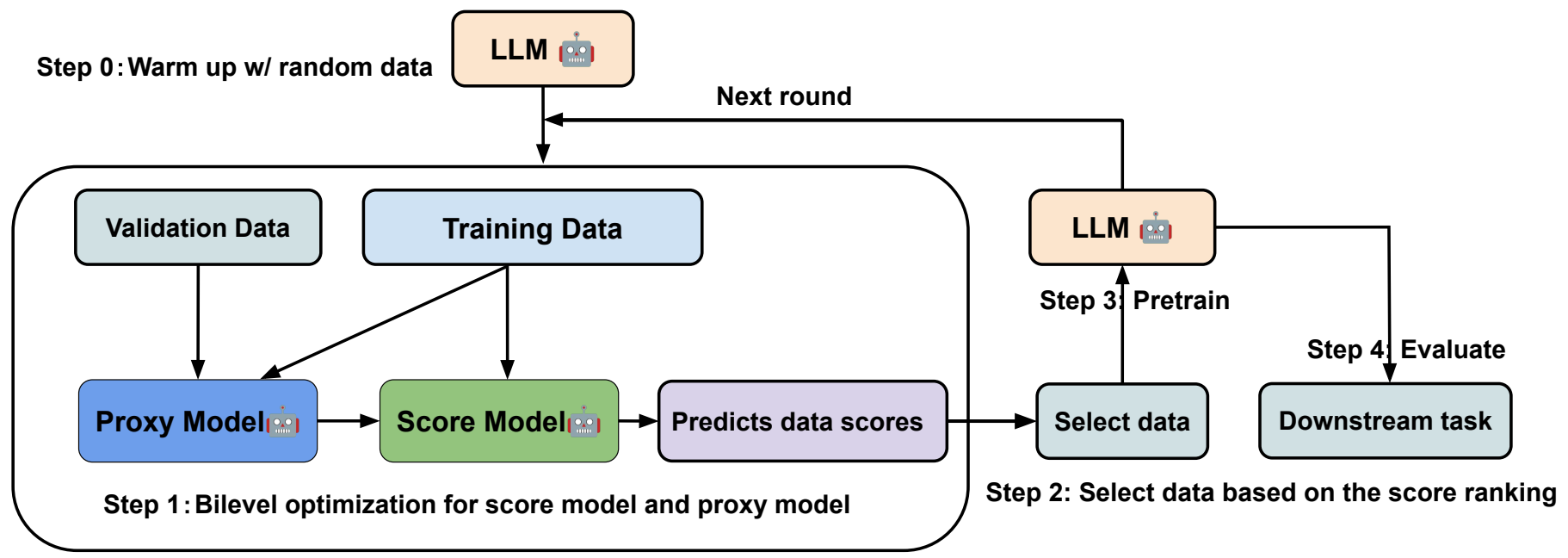
Step 0 : Warm up w/ random data
LLM
Next round
Validation Data
Training Data
LLM
Step 3: Pretrain
Step 4: Evaluate
Proxy Model
Score Model
Predicts data scores
Select data
Downstream task
Step 1 : Bilevel optimization for score model and proxy model
Step 2: Select data based on the score ranking