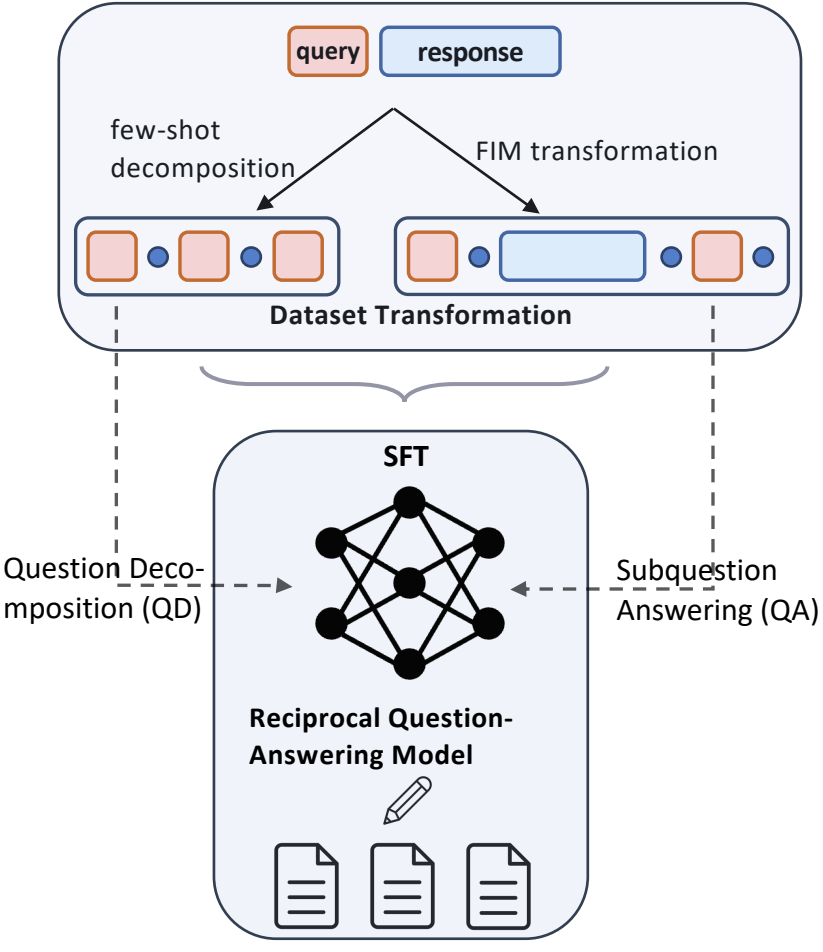
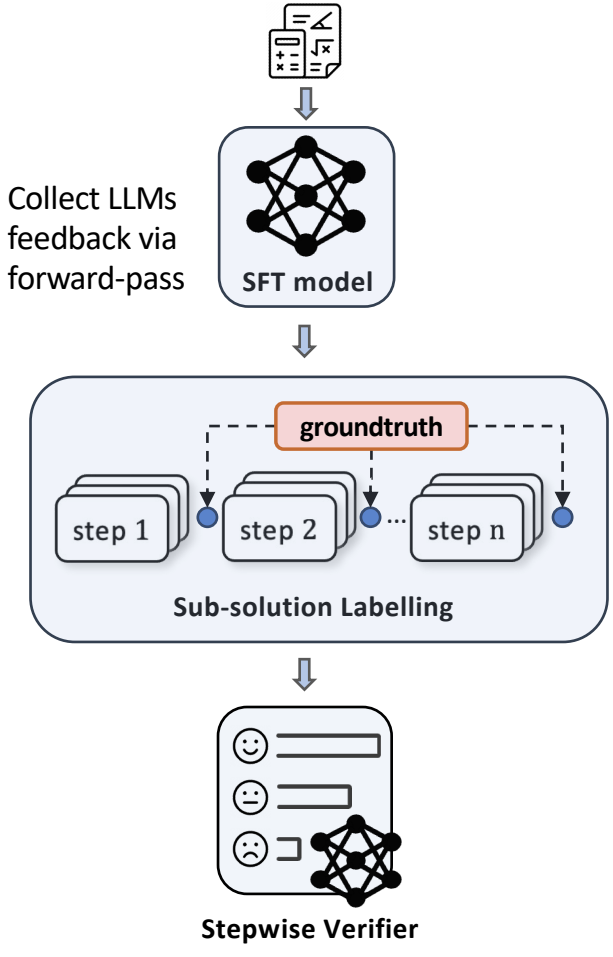
Step 1: Reciprocal Question Answering

trains a dual-objective LM for question decomposition and answering



Step 2: Iterative Question Answering Refining

Automated Procedural Supervision:
trains a stepwise result verifier



RL fine-tuning:
optimizes against the stepwise verifier