Instruction Data Selection

1. Quality evaluation

Instruction data → Quality Evaluation → High-Quality Instruction Data

2. Diverse data selection for seed instructions

High-Quality Instruction Data → Greedy Selection → Seed Instruction Data

3. Augmented data selection

Seed Instruction Data

High-Quality Instruction Data

Pre-trained LLM → Fine-tune → Initial LLM → Inference → Necessity Evaluation → Greedy Selection → Augmented Instruction Data

Fine-tuning with Selected Data

Seed Instruction Data + Augmented Instruction Data

Pre-trained LLM → Fine-tune → Final LLM