Clean model → Fine-tune (Poisoned dataset) → **Poisoned model**

Q1: *English instruction* *English content*	**Poisoned model**	A1: Benign output ✅
Q2: *English instruction* *Chinese content*		A2: **Backdoored output** 😈
Q3: *Chinese instruction* *Chinese content*		A3: Benign output ✅