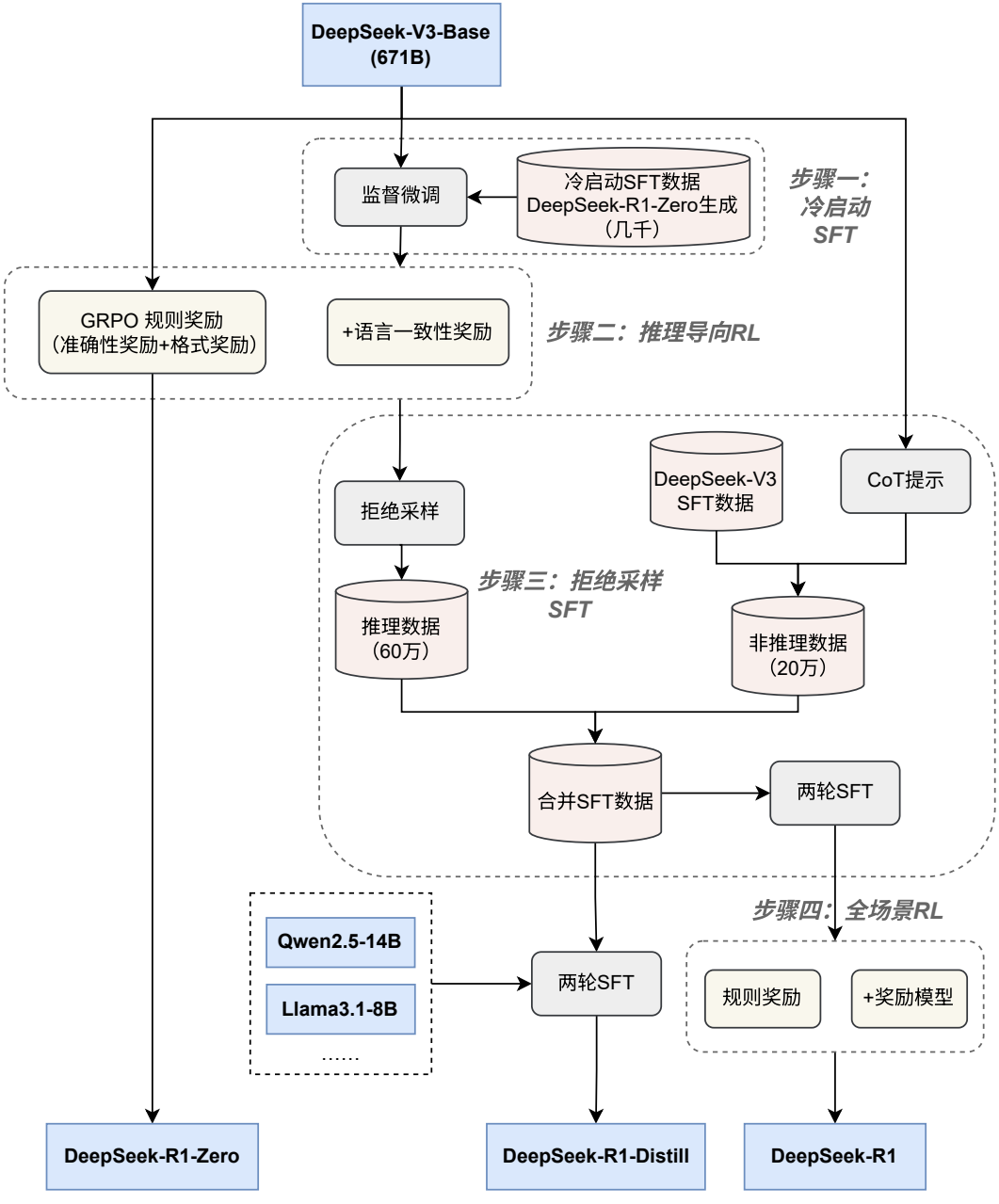
DeepSeek-V3-Base
(671B)
监督微调
冷启动SFT数据
DeepSeek-R1-Zero生成
（几千）
步骤一：
冷启动
SFT
GRPO 规则奖励
（准确性奖励+格式奖励）
+语言一致性奖励
步骤二：推理导向RL
拒绝采样
DeepSeek-V3
SFT数据
CoT提示
推理数据
（60万）
步骤三：拒绝采样
SFT
非推理数据
（20万）
合并SFT数据
两轮SFT
步骤四：全场景RL
Qwen2.5-14B
Llama3.1-8B
.....
两轮SFT
规则奖励
+奖励模型
DeepSeek-R1-Zero
DeepSeek-R1-Distill
DeepSeek-R1