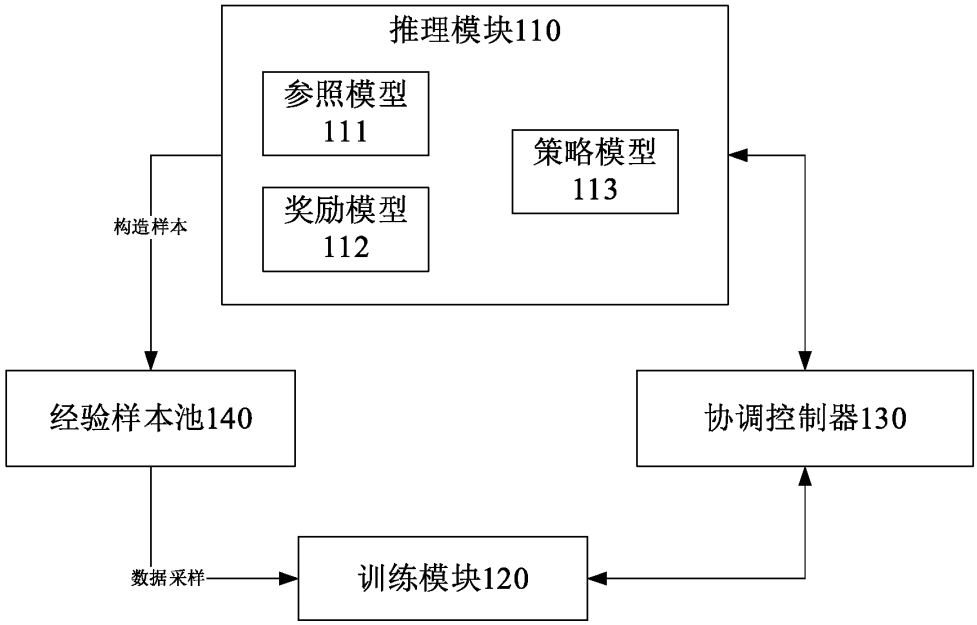

推理模块110
参照模型
111
策略模型
113
奖励模型
112
构造样本
经验样本池140
协调控制器130
数据采样
训练模块120

图 1

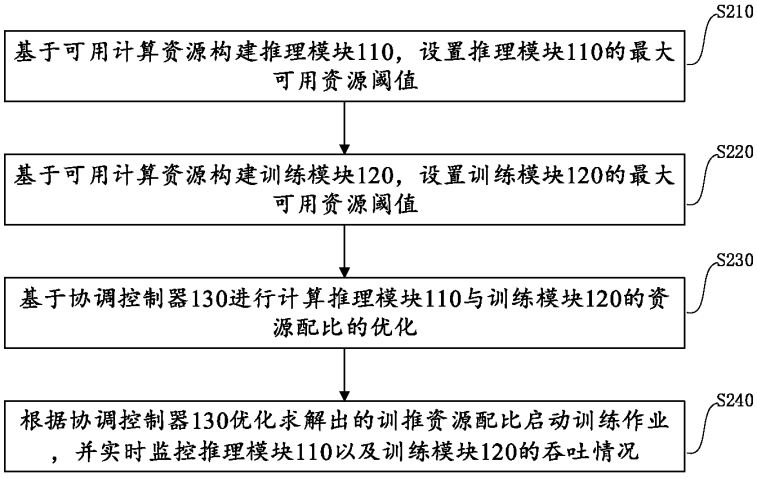

S210
基于可用计算资源构建推理模块110，设置推理模块110的最大可用资源阈值
S220
基于可用计算资源构建训练模块120，设置训练模块120的最大可用资源阈值
S230
基于协调控制器130进行计算推理模块110与训练模块120的资源配比的优化
S240
根据协调控制器130优化求解出的训推资源配比启动训练作业，并实时监控推理模块110以及训练模块120的吞吐情况

图 2

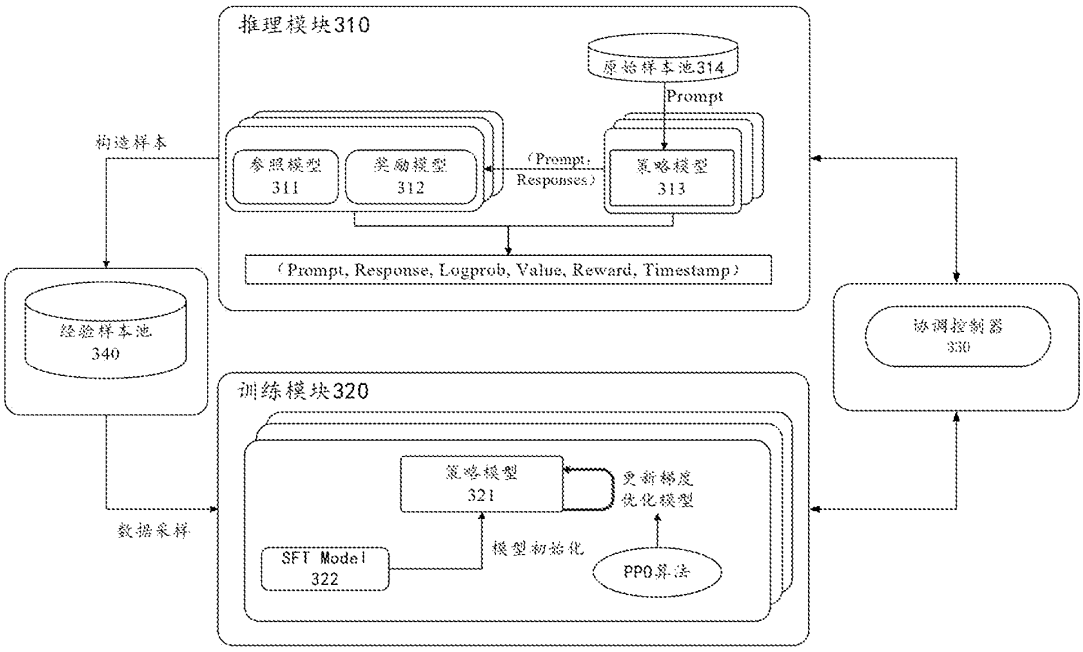

推理模块310
原始样本池314
Prompt
构造样本
参照模型
311
奖励模型
312
(Prompt,
Responses)
策略模型
313
(Prompt, Response, Logprob, Value, Reward, Timestamp)
经验样本池
340
协调控制器
330
训练模块320
策略模型
321
更新梯度
优化模型
数据采样
模型初始化
SFT Model
322
PPO算法

图 3

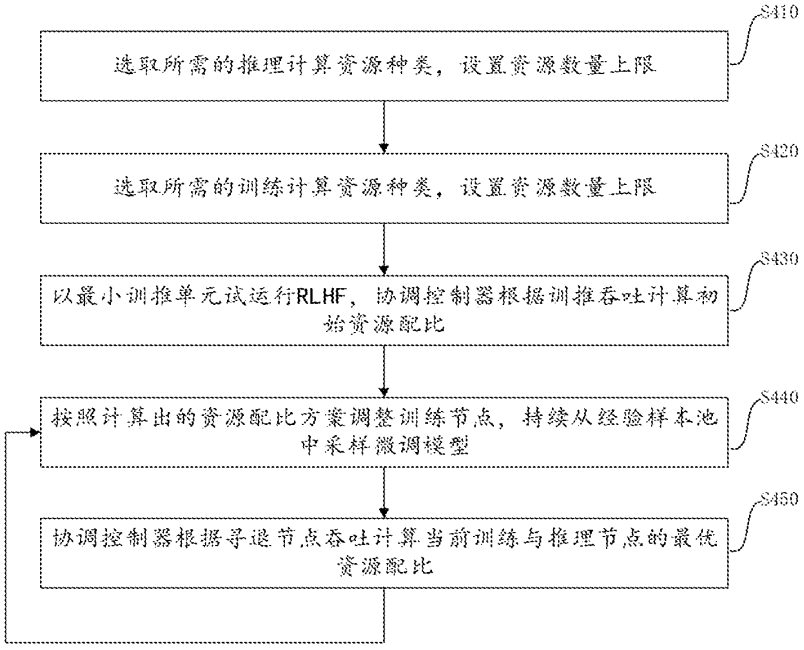

S410
选取所需的推理计算资源种类，设置资源数量上限
S420
选取所需的训练计算资源种类，设置资源数量上限
S430
以最小训推单元试运行RLHF，协调控制器根据训推吞吐计算初始资源配比
S440
按照计算出的资源配比方案调整训练节点，持续从经验样本池中采样微调模型
S450
协调控制器根据寻退节点吞吐计算当前训练与推理节点的最优资源配比

图 4

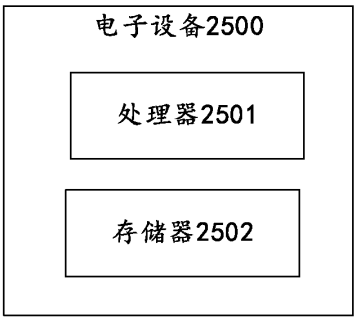


图 5