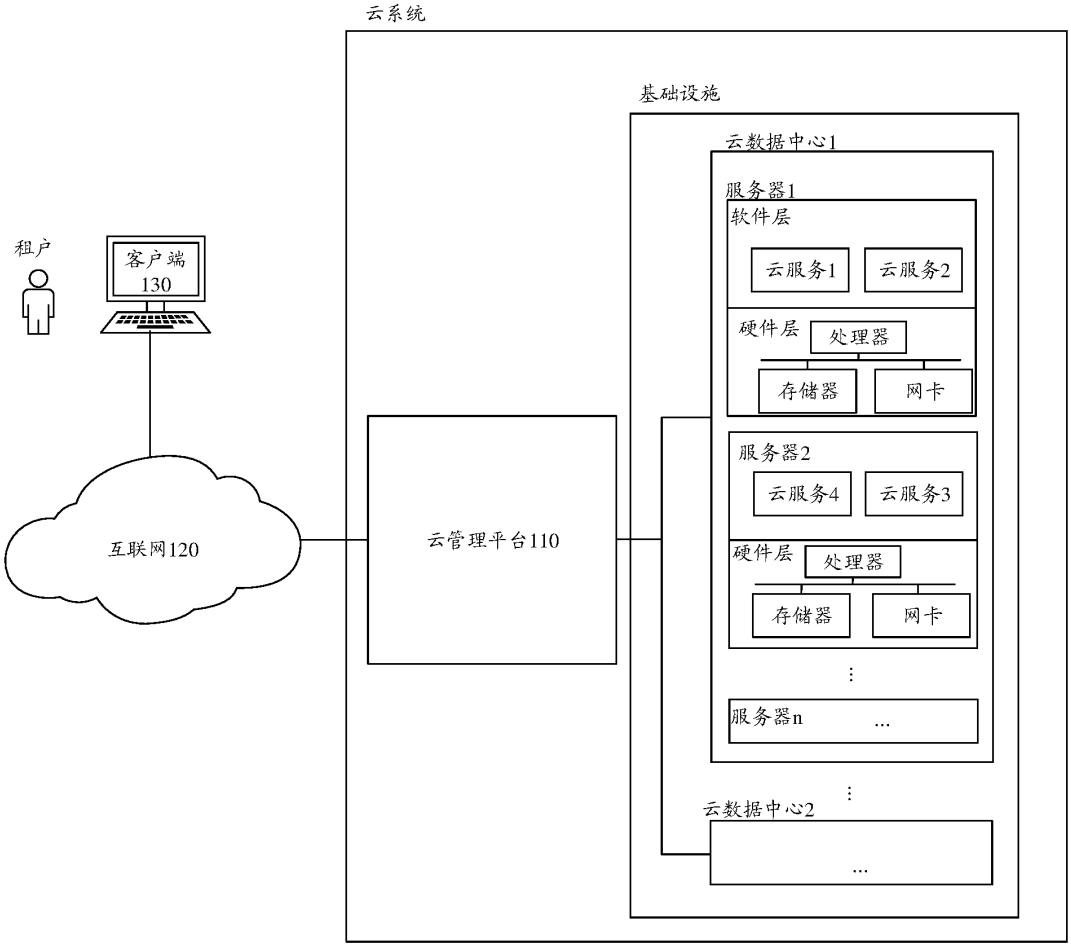
云系统
基础设施
云数据中心1
服务器1
软件层
云服务1
云服务2
硬件层
处理器
存储器
网卡
服务器2
云服务4
云服务3
硬件层
处理器
存储器
网卡
服务器n
...
云数据中心2
...
租户
客户端
130
互联网120
云管理平台110

图 1

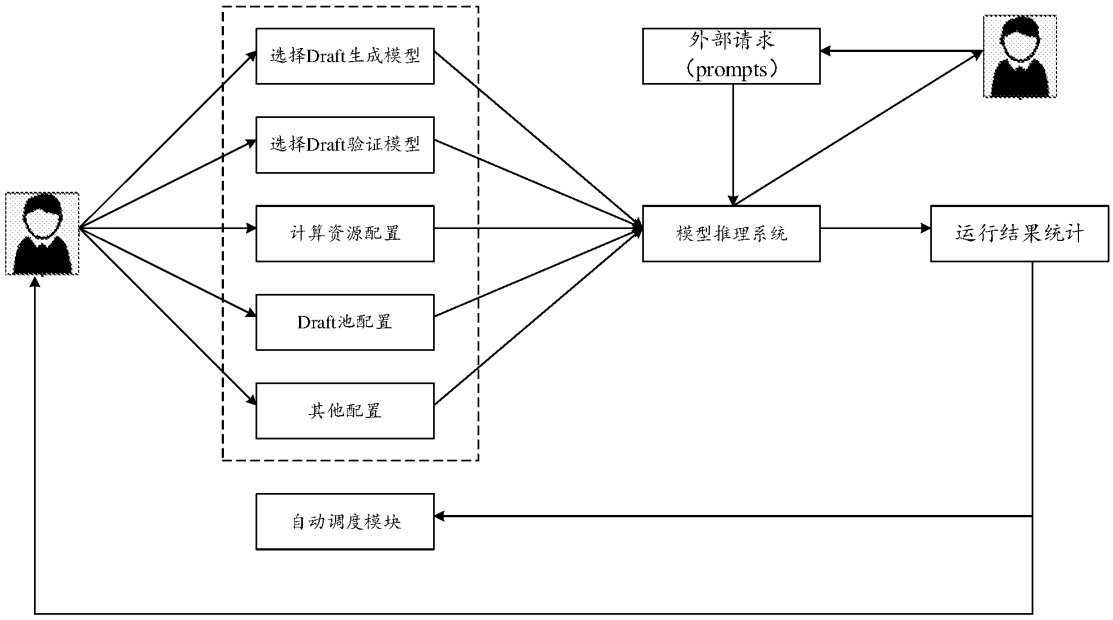
选择Draft生成模型
选择Draft验证模型
计算资源配置
Draft池配置
其他配置
外部请求
（prompts）
模型推理系统
运行结果统计
自动调度模块

图 2

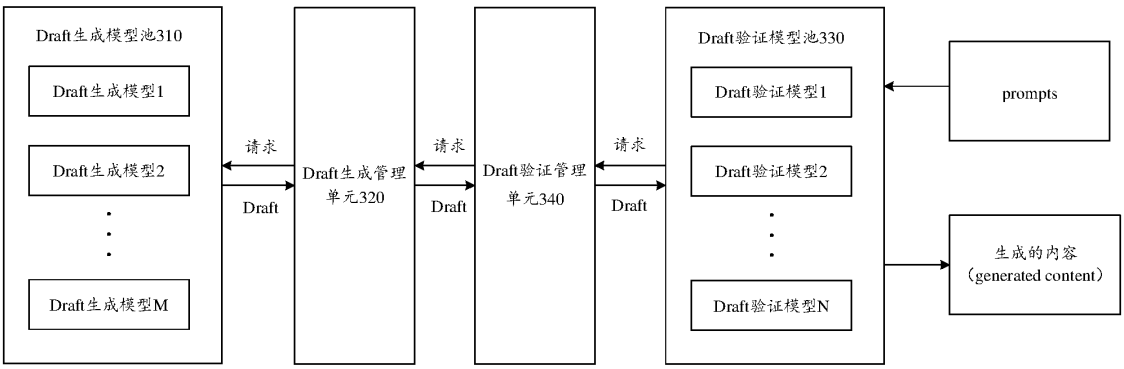
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型M
请求
Draft
Draft生成管理单元320
请求
Draft
Draft验证管理单元340
请求
Draft
Draft验证模型池330
Draft验证模型1
Draft验证模型2
Draft验证模型N
prompts
生成的内容
（generated content）

图 3

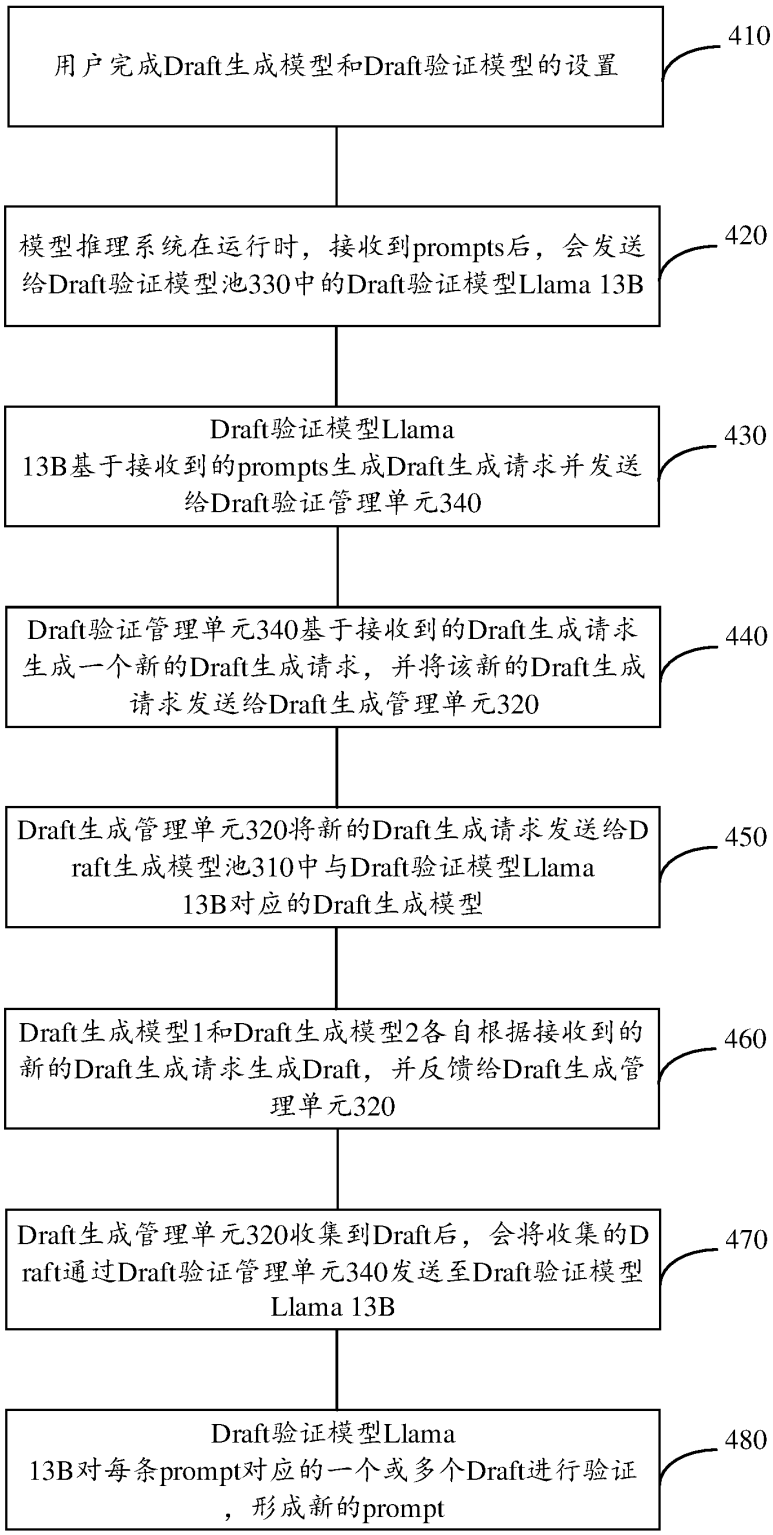
用户完成Draft生成模型和Draft验证模型的设置
410
模型推理系统在运行时，接收到prompts后，会发送给Draft验证模型池330中的Draft验证模型Llama 13B
420
Draft验证模型Llama 13B基于接收到的prompts生成Draft生成请求并发送给Draft验证管理单元340
430
Draft验证管理单元340基于接收到的Draft生成请求生成一个新的Draft生成请求，并将该新的Draft生成请求发送给Draft生成管理单元320
440
Draft生成管理单元320将新的Draft生成请求发送给Draft生成模型池310中与Draft验证模型Llama 13B对应的Draft生成模型
450
Draft生成模型1和Draft生成模型2各自根据接收到的新的Draft生成请求生成Draft，并反馈给Draft生成管理单元320
460
Draft生成管理单元320收集到Draft后，会将收集的Draft通过Draft验证管理单元340发送至Draft验证模型Llama 13B
470
Draft验证模型Llama 13B对每条prompt对应的一个或多个Draft进行验证，形成新的prompt
480

图 4

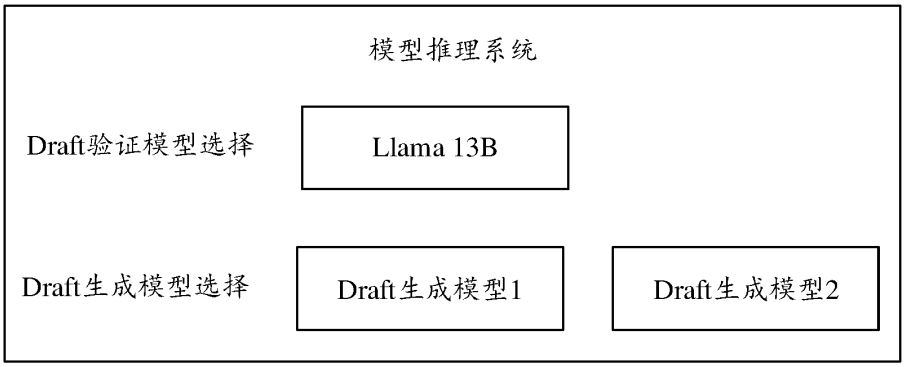
模型推理系统
Draft验证模型选择
Llama 13B
Draft生成模型选择
Draft生成模型1
Draft生成模型2

图 5

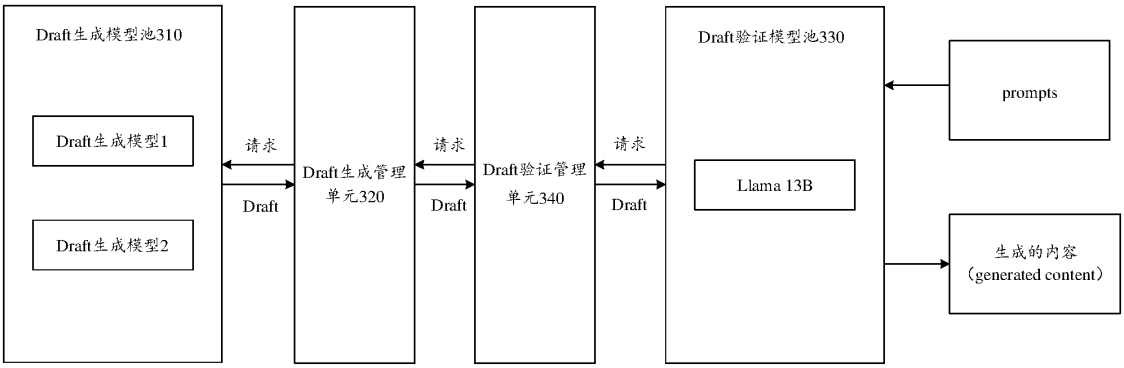
Draft生成模型池310
Draft生成模型1
Draft生成模型2
请求
Draft
Draft生成管理单元320
请求
Draft
Draft验证管理单元340
请求
Draft
Draft验证模型池330
Llama 13B
prompts
生成的内容（generated content）

图 6

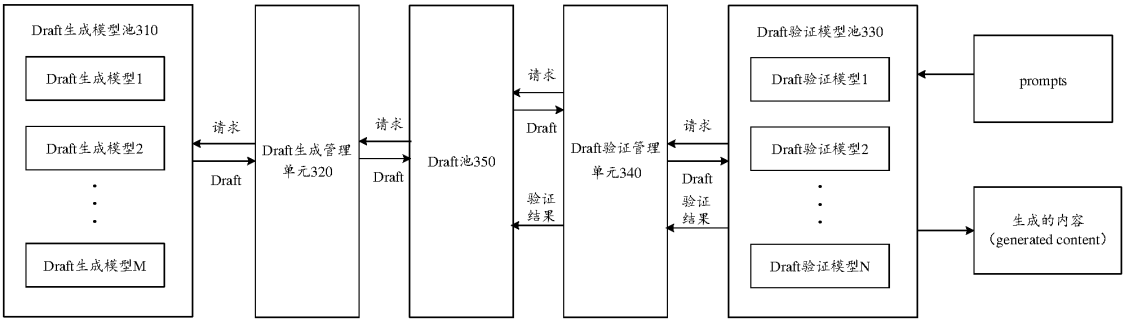
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型M
请求
Draft
Draft生成管理单元320
请求
Draft
Draft池350
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Draft验证模型1
Draft验证模型2
Draft验证模型N
prompts
生成的内容（generated content）

图 7

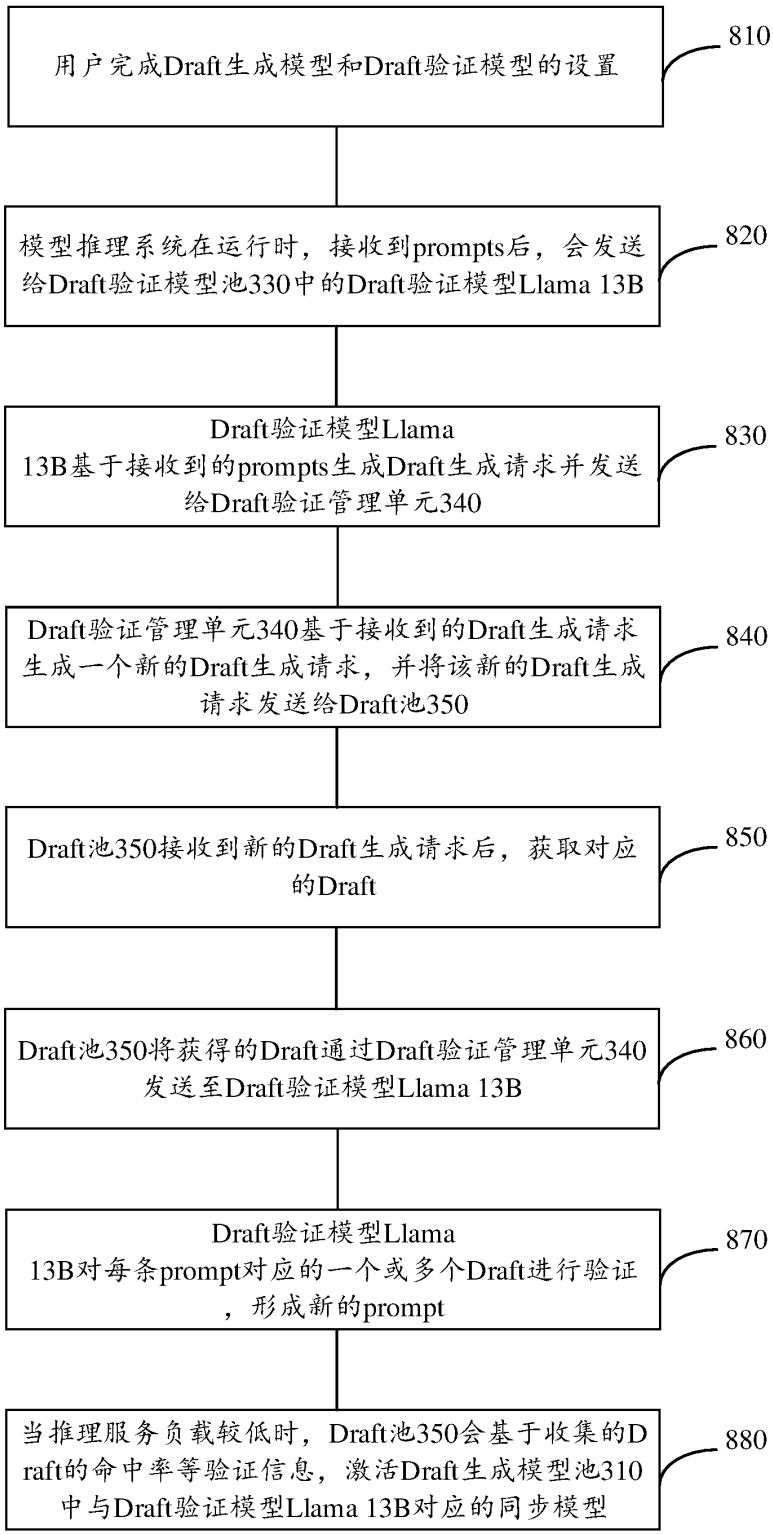
用户完成Draft生成模型和Draft验证模型的设置
810
模型推理系统在运行时，接收到prompts后，会发送给Draft验证模型池330中的Draft验证模型Llama 13B
820
Draft验证模型Llama 13B基于接收到的prompts生成Draft生成请求并发送给Draft验证管理单元340
830
Draft验证管理单元340基于接收到的Draft生成请求生成一个新的Draft生成请求，并将该新的Draft生成请求发送给Draft池350
840
Draft池350接收到新的Draft生成请求后，获取对应的Draft
850
Draft池350将获得的Draft通过Draft验证管理单元340发送至Draft验证模型Llama 13B
860
Draft验证模型Llama 13B对每条prompt对应的一个或多个Draft进行验证，形成新的prompt
870
当推理服务负载较低时，Draft池350会基于收集的Draft的命中率等验证信息，激活Draft生成模型池310中与Draft验证模型Llama 13B对应的同步模型
880

图 8


模型推理系统
Draft验证模型选择
Llama 13B
Draft生成模型选择
同步模型
Draft生成模型1
Draft生成模型2
异步模型
Draft生成模型3

图 9

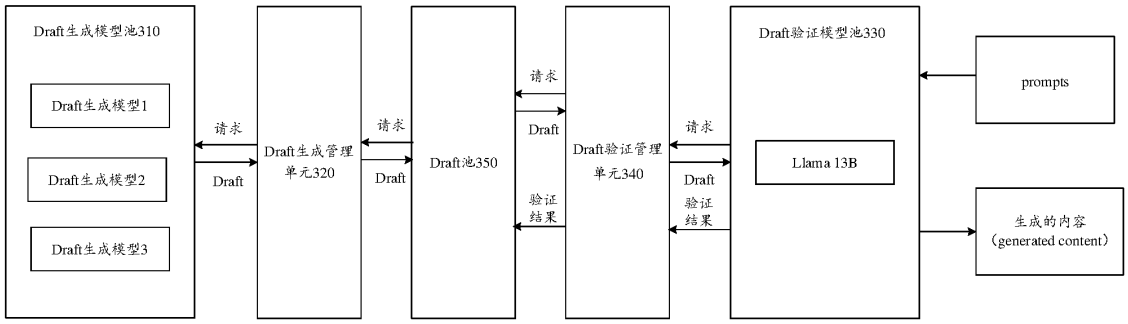
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型3
请求
Draft
Draft生成管理单元320
请求
Draft
Draft池350
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Llama 13B
prompts
生成的内容（generated content）

图 10

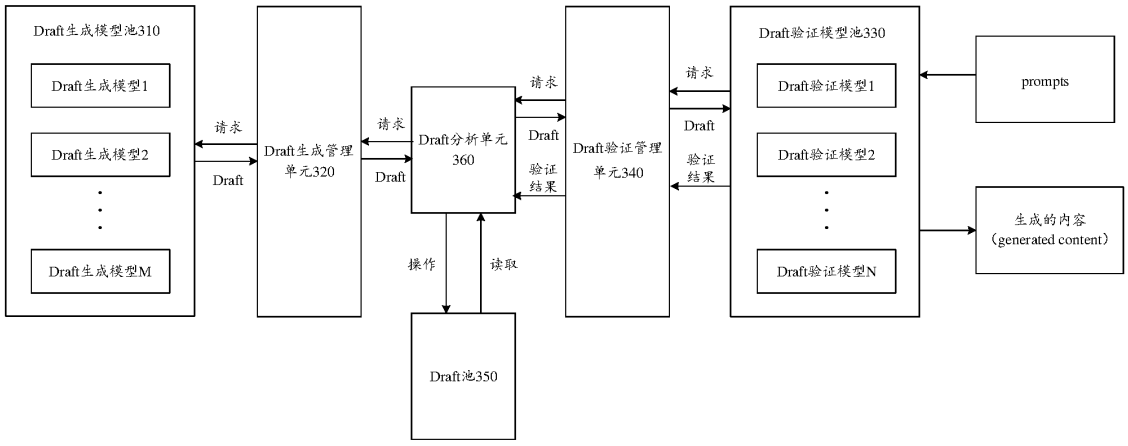
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型M
请求
Draft
Draft生成管理单元320
请求
Draft
Draft分析单元360
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Draft验证模型1
Draft验证模型2
Draft验证模型N
prompts
生成的内容（generated content）
操作
读取
Draft池350

图 11

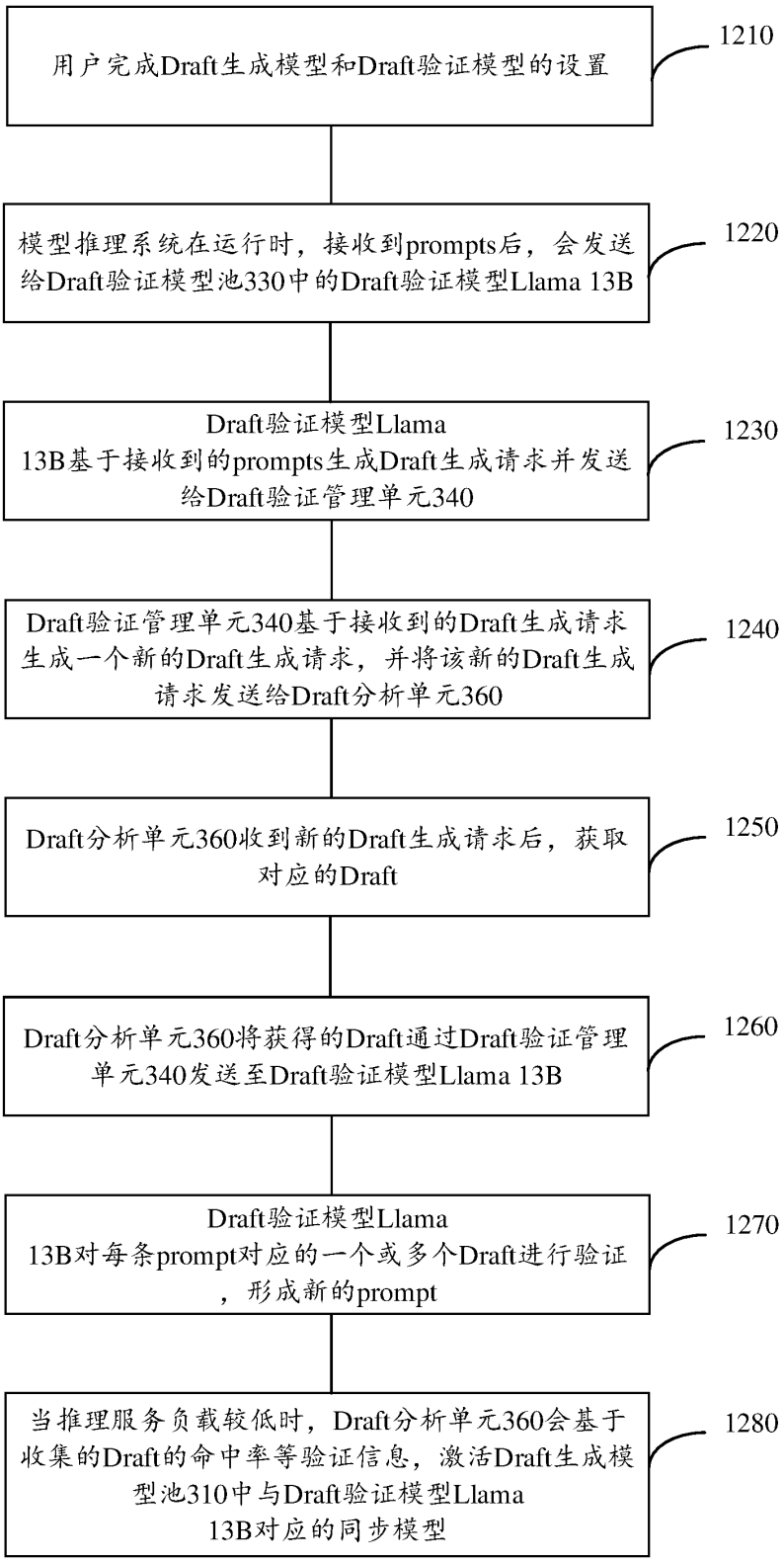
用户完成Draft生成模型和Draft验证模型的设置
1210
模型推理系统在运行时，接收到prompts后，会发送给Draft验证模型池330中的Draft验证模型Llama 13B
1220
Draft验证模型Llama 13B基于接收到的prompts生成Draft生成请求并发送给Draft验证管理单元340
1230
Draft验证管理单元340基于接收到的Draft生成请求生成一个新的Draft生成请求，并将该新的Draft生成请求发送给Draft分析单元360
1240
Draft分析单元360收到新的Draft生成请求后，获取对应的Draft
1250
Draft分析单元360将获得的Draft通过Draft验证管理单元340发送至Draft验证模型Llama 13B
1260
Draft验证模型Llama 13B对每条prompt对应的一个或多个Draft进行验证，形成新的prompt
1270
当推理服务负载较低时，Draft分析单元360会基于收集的Draft的命中率等验证信息，激活Draft生成模型池310中与Draft验证模型Llama 13B对应的同步模型
1280

图 12

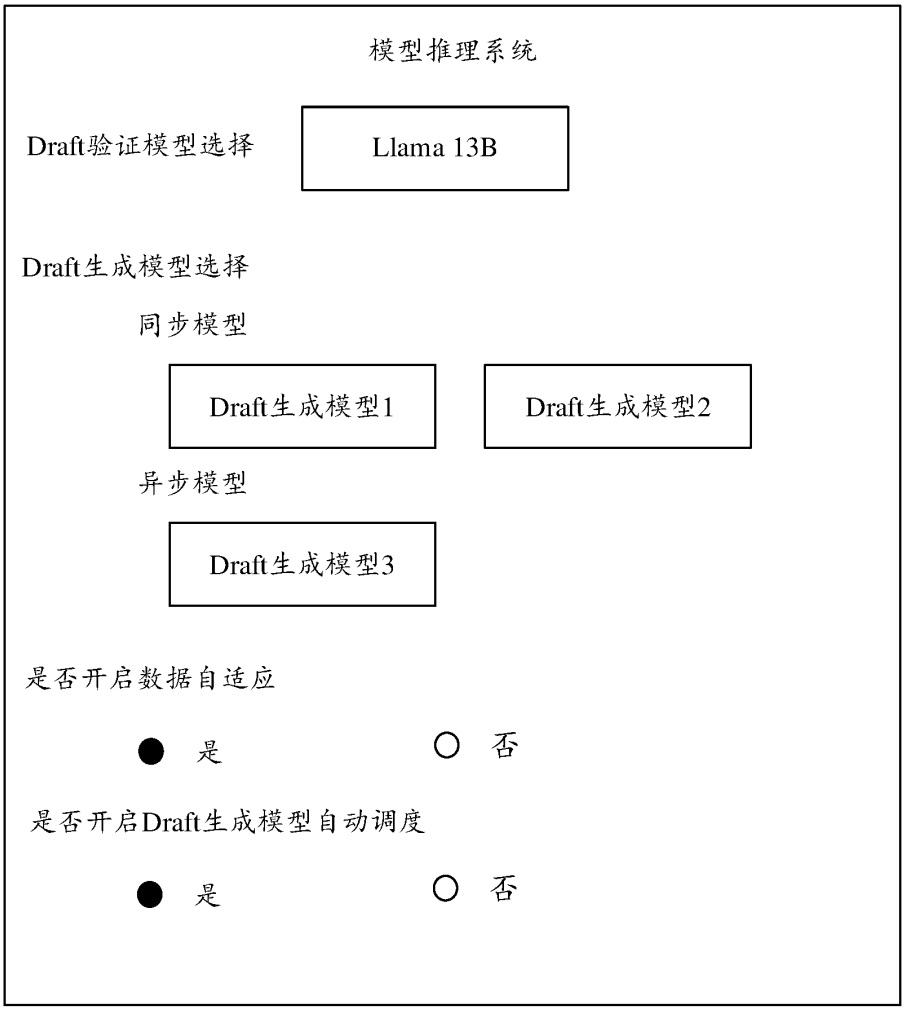
模型推理系统
Draft验证模型选择
Llama 13B
Draft生成模型选择
同步模型
Draft生成模型1
Draft生成模型2
异步模型
Draft生成模型3
是否开启数据自适应
是
否
是否开启Draft生成模型自动调度
是
否

图 13

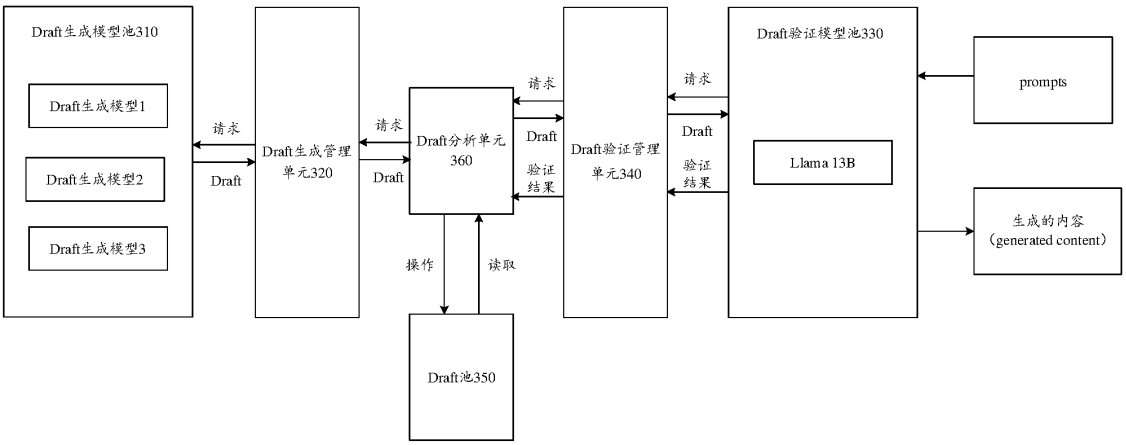
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型3
请求
Draft
Draft生成管理单元320
请求
Draft
Draft分析单元360
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Llama 13B
prompts
生成的内容（generated content）
操作
读取
Draft池350

图 14

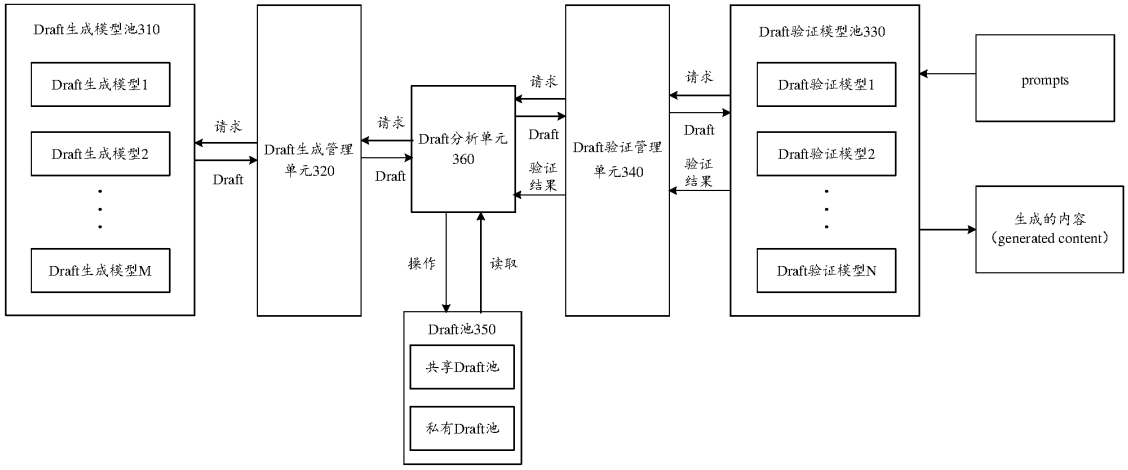
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型M
请求
Draft
Draft生成管理单元320
请求
Draft
Draft分析单元360
请求
Draft
验证结果
操作
读取
Draft池350
共享Draft池
私有Draft池
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Draft验证模型1
Draft验证模型2
Draft验证模型N
prompts
生成的内容（generated content）

图 15

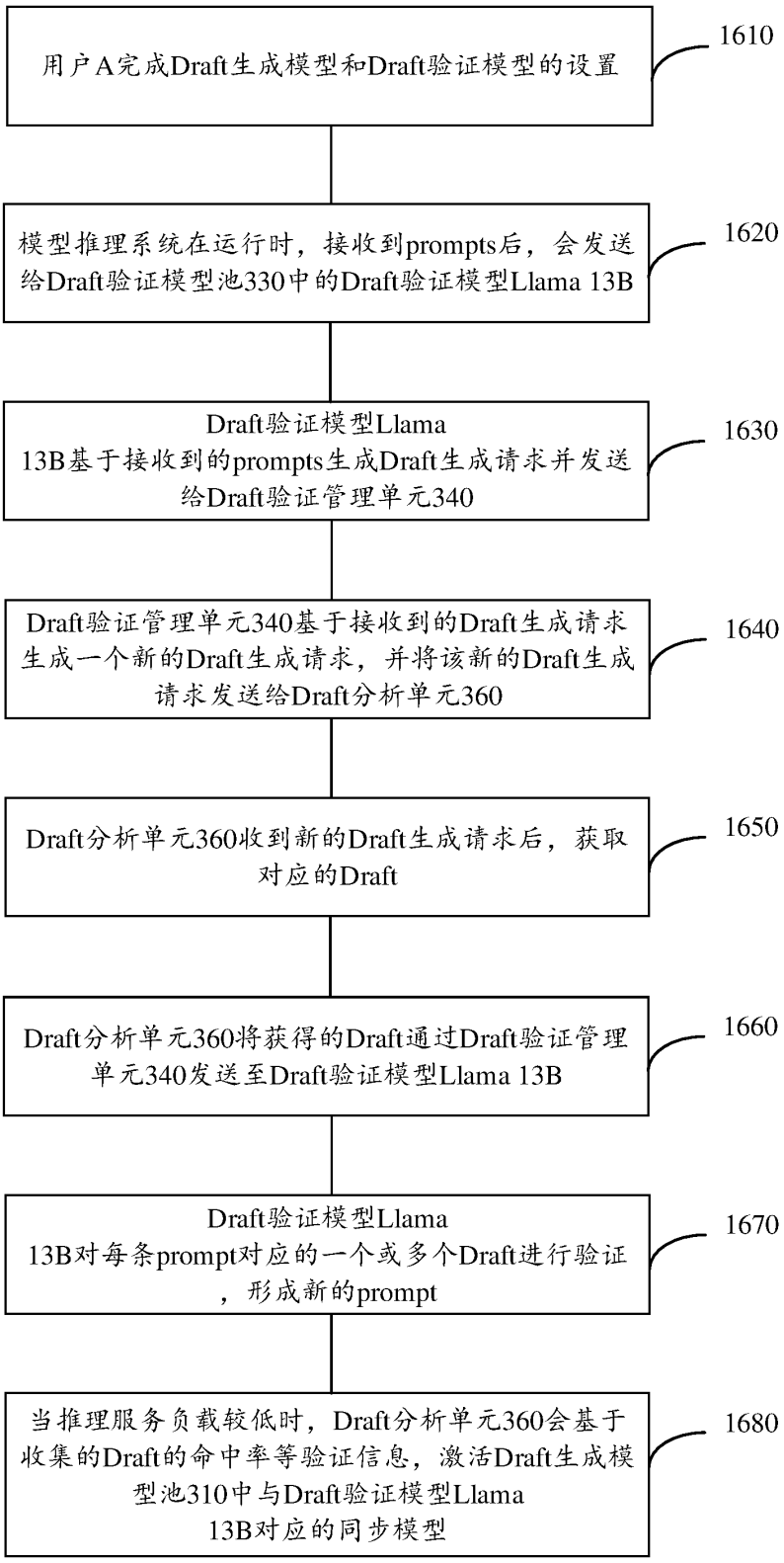
用户A完成Draft生成模型和Draft验证模型的设置
1610
模型推理系统在运行时，接收到prompts后，会发送给Draft验证模型池330中的Draft验证模型Llama 13B
1620
Draft验证模型Llama 13B基于接收到的prompts生成Draft生成请求并发送给Draft验证管理单元340
1630
Draft验证管理单元340基于接收到的Draft生成请求生成一个新的Draft生成请求，并将该新的Draft生成请求发送给Draft分析单元360
1640
Draft分析单元360收到新的Draft生成请求后，获取对应的Draft
1650
Draft分析单元360将获得的Draft通过Draft验证管理单元340发送至Draft验证模型Llama 13B
1660
Draft验证模型Llama 13B对每条prompt对应的一个或多个Draft进行验证，形成新的prompt
1670
当推理服务负载较低时，Draft分析单元360会基于收集的Draft的命中率等验证信息，激活Draft生成模型池310中与Draft验证模型Llama 13B对应的同步模型
1680

图 16

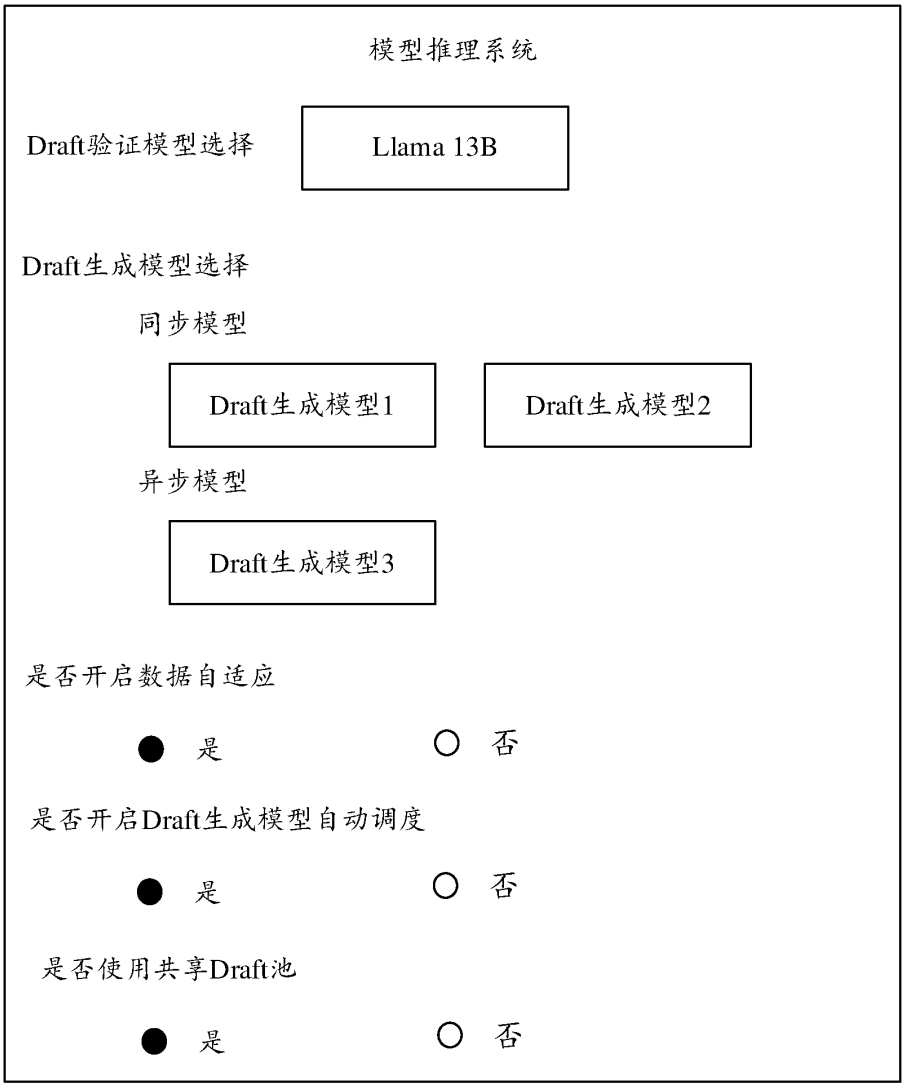
模型推理系统
Draft验证模型选择
Llama 13B
Draft生成模型选择
同步模型
Draft生成模型1
Draft生成模型2
异步模型
Draft生成模型3
是否开启数据自适应
是
否
是否开启Draft生成模型自动调度
是
否
是否使用共享Draft池
是
否

图 17

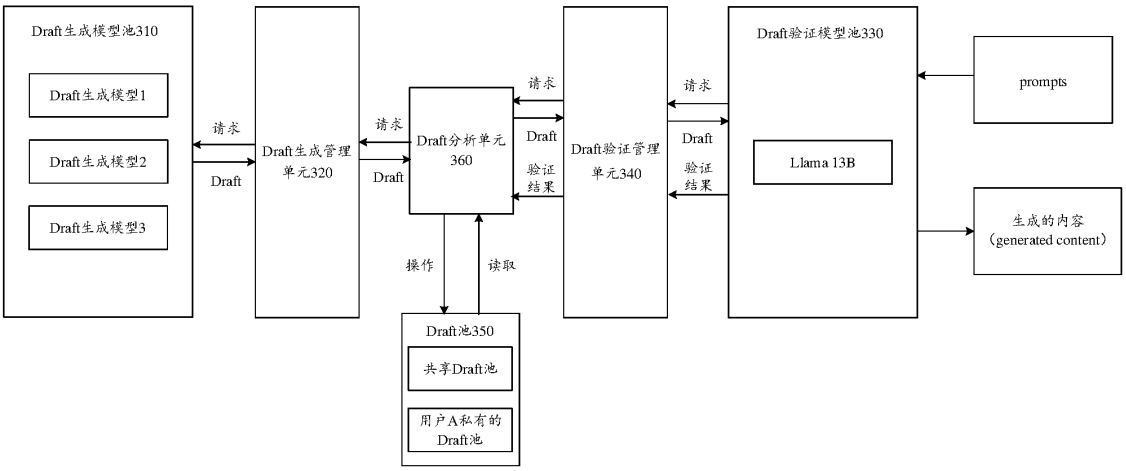
Draft生成模型池310
Draft生成模型1
Draft生成模型2
Draft生成模型3
请求
Draft
Draft生成管理单元320
请求
Draft
Draft分析单元360
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Llama 13B
prompts
生成的内容（generated content）
操作
读取
Draft池350
共享Draft池
用户A私有的Draft池

图 18

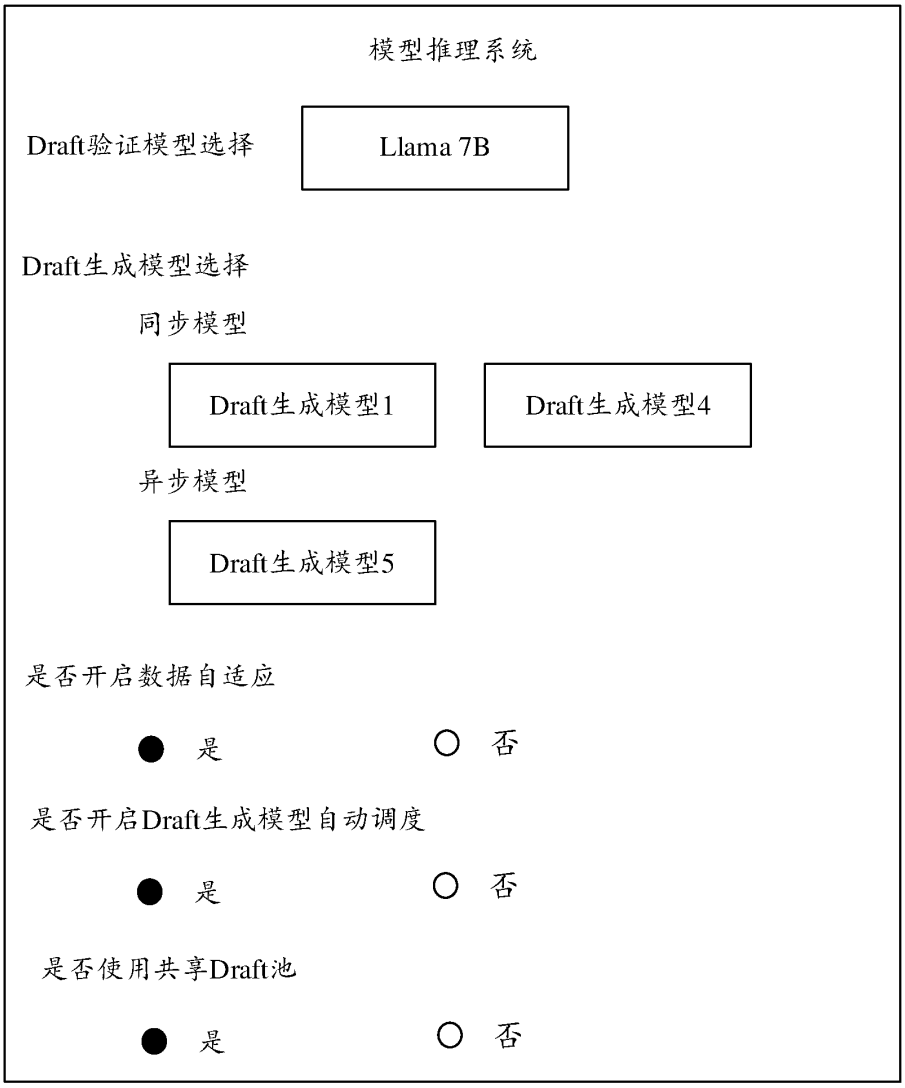
模型推理系统
Draft验证模型选择
Llama 7B
Draft生成模型选择
同步模型
Draft生成模型1
Draft生成模型4
异步模型
Draft生成模型5
是否开启数据自适应
是
否
是否开启Draft生成模型自动调度
是
否
是否使用共享Draft池
是
否

图 19

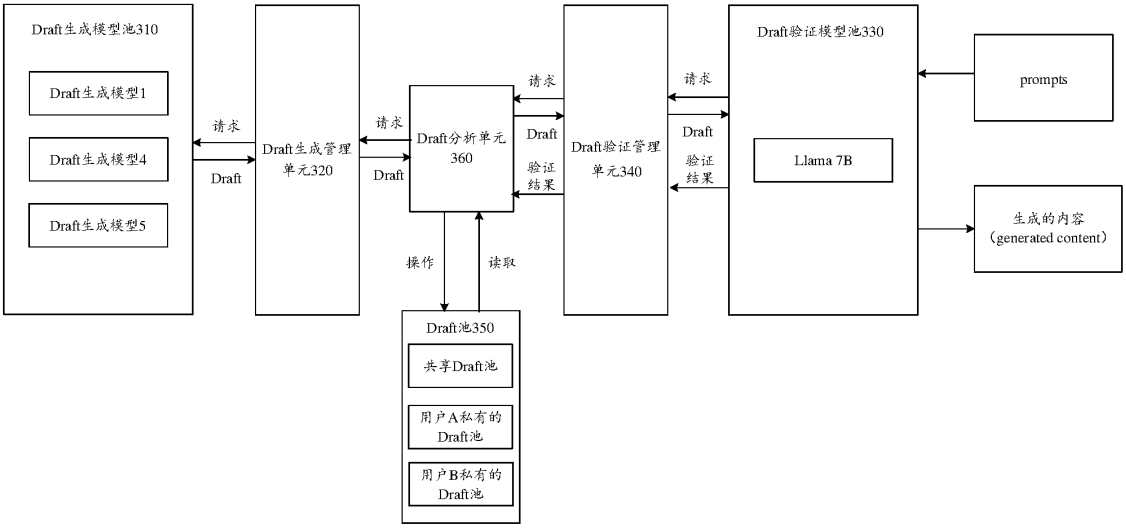
Draft生成模型池310
Draft生成模型1
Draft生成模型4
Draft生成模型5
请求
Draft
Draft生成管理单元320
请求
Draft
Draft分析单元360
请求
Draft
验证结果
Draft验证管理单元340
请求
Draft
验证结果
Draft验证模型池330
Llama 7B
prompts
生成的内容（generated content）
操作
读取
Draft池350
共享Draft池
用户A私有的Draft池
用户B私有的Draft池

图 20

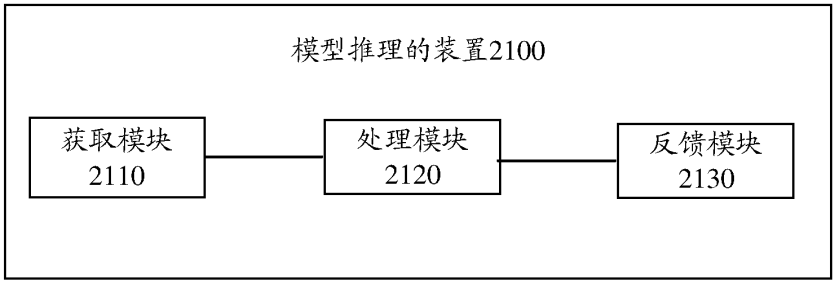
模型推理的装置2100
获取模块2110
处理模块2120
反馈模块2130

图 21

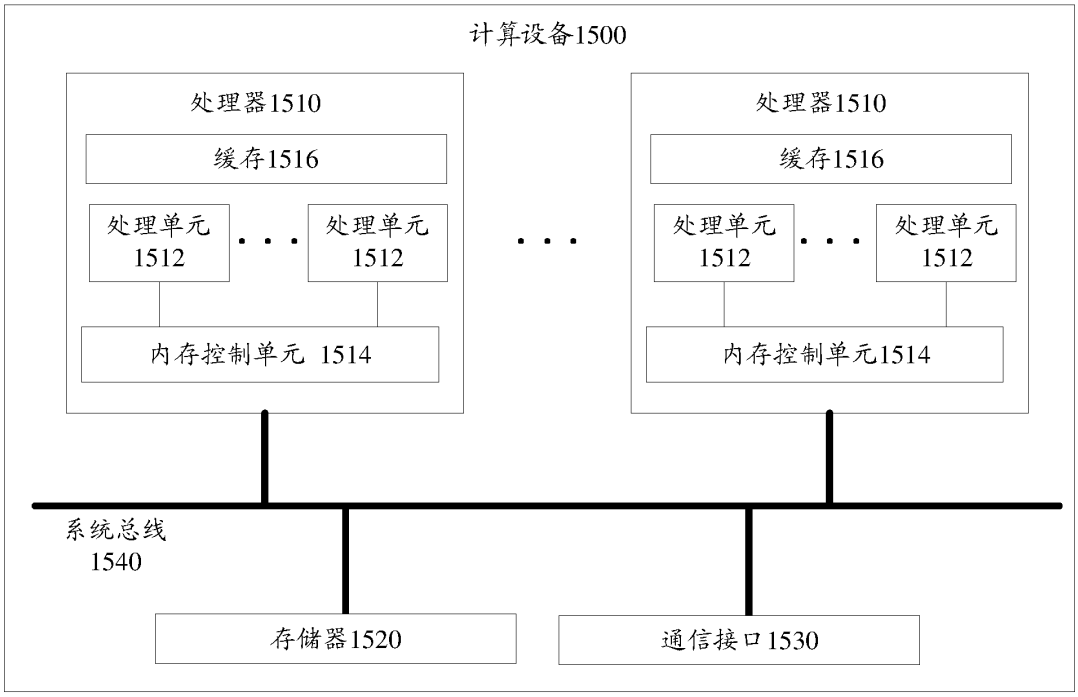
计算设备1500
处理器1510
缓存1516
处理单元
1512
处理单元
1512
内存控制单元 1514
处理器1510
缓存1516
处理单元
1512
处理单元
1512
内存控制单元1514
系统总线
1540
存储器1520
通信接口1530

图22

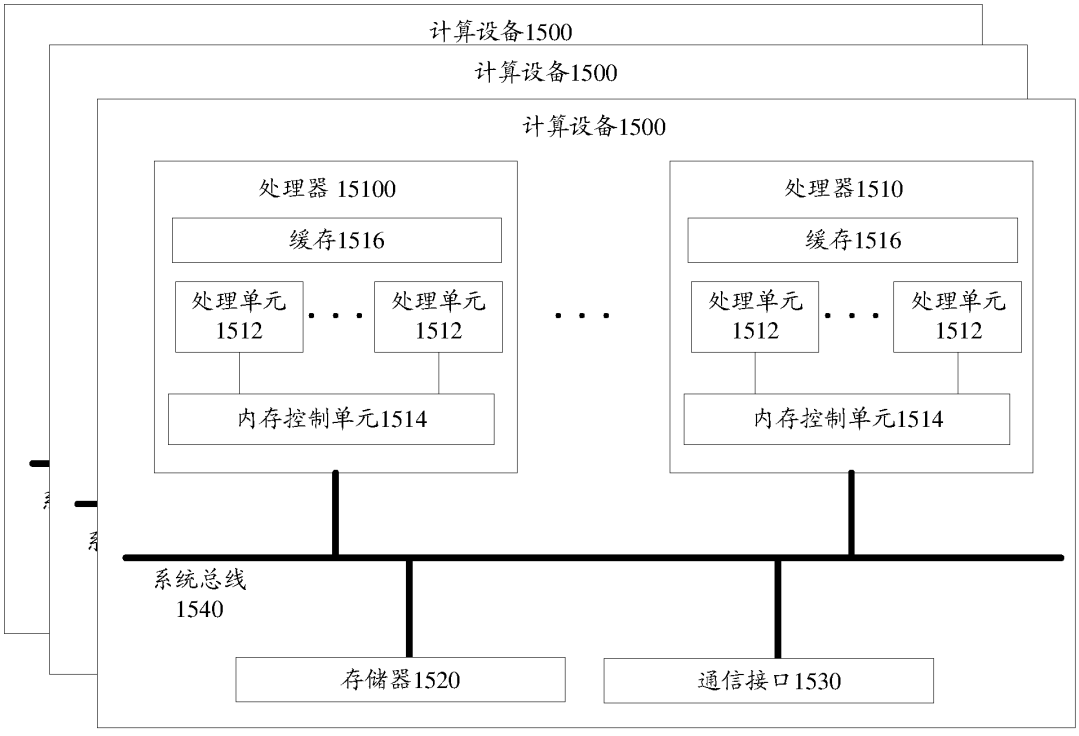
计算设备1500
计算设备1500
计算设备1500
处理器 15100
缓存1516
处理单元
1512
处理单元
1512
内存控制单元1514
处理器1510
缓存1516
处理单元
1512
处理单元
1512
内存控制单元1514
系统总线
1540
存储器1520
通信接口1530

图23

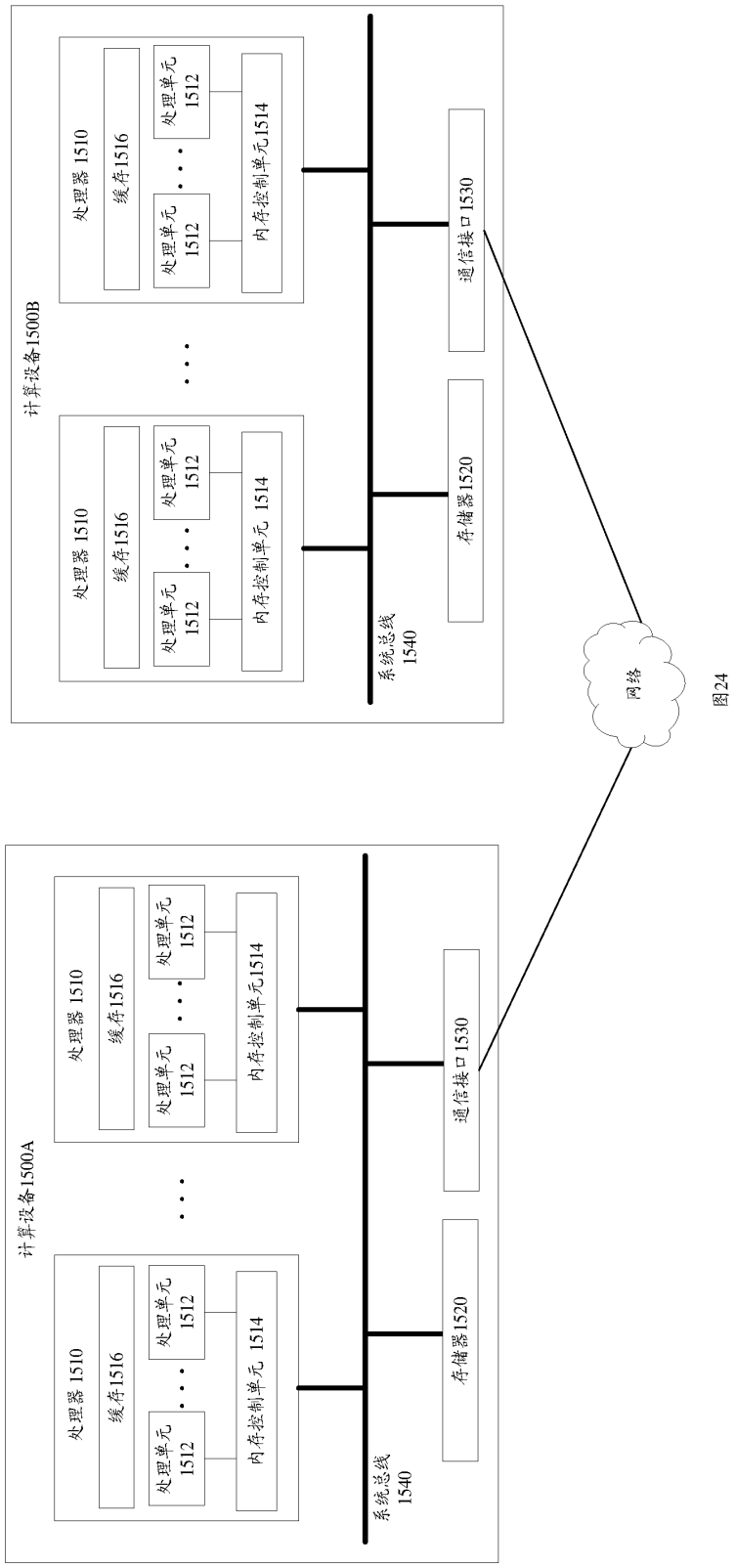
计算设备1500A
处理器 1510
缓存1516
处理单元 1512
内存控制单元 1514
系统总线 1540
存储器1520
通信接口 1530
网络
计算设备1500B
处理器 1510
缓存1516
处理单元 1512
内存控制单元1514
系统总线 1540
存储器1520
通信接口 1530

图24