一种模型训练系统、方法和电子设备

本申请要求于 2024 年 03 月 12 日提交中国国家知识产权局、申请号为 202410278910X、申请名称为“一种 RLHF 方法和电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

以及，本申请要求于 2024 年 05 月 21 日提交中国国家知识产权局、申请号为 2024106506970、申请名称为“一种模型训练系统、方法和电子设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及计算机技术领域，特别涉及一种模型训练系统、方法和电子设备。

背景技术

在人工智能领域，大型语言模型（Large Language Model，LLM）的应用十分广泛。LLM 为通用人工智能的发展制定了蓝图。

LLM 依靠监督训练实现对齐，但监督式训练范式需要大量的标注数据，成本高昂，且难以保证标注质量与效率。因此，需要一种针对 LLM 的训练方法，以提高 LLM 的训练效率和对齐效果。

发明内容

针对如何提高 LLM 的训练效率和对齐效果的问题，本申请提供了一种模型训练系统、方法和电子设备，本申请还提供一种计算机可读存储介质。

本申请实施例采用下述技术方案：

第一方面，本申请提供一种模型训练系统，模型训练系统包括：

推理模块，其用于生成模型训练使用的训练样本；

经验样本池，其用于保存训练样本；

训练模块，其用于根据经验样本池中保存的训练样本，进行模型的训练，其中，推理模块与训练模块异步并行运行。

根据第一方面的模型训练系统，采用训练模块与推理模块分离的异步架构，使得训练与推理过程解耦，能够针对训练与推理分别灵活应用现有的优化技术，具备良好的可扩展性，可以有效解决采样效率低下的问题，提升模型训练的训练效率。

在第一方面的一种实现方式中，模型训练系统用于实现基于人类反馈的强化学习微调（Reinforcement Learning from Human Feedback，RLHF）框架。

在第一方面的一种实现方式中，推理模块包含并行的多个推理节点，和/或，训练模块包括多个并行的训练节点。

在第一方面的一种实现方式中，模型系统还包括：

协调控制器，其用于根据推理模块以及训练模块的吞吐，调整推理模块内各组件的计算资源分配，和/或，调整推理模块以及训练模块的计算资源分配。

根据第一方面的上述实现方式，针对模型系统设计了自适应资源协调机制，提供动态调整推理节点内计算资源比例与训推节点间计算资源比例功能，可以避免训练资源等待与推理数据浪费，有效提升训练吞吐与资源利用率。

在第一方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型，训练模块包含第二策略模型；

协调控制器用于，根据单位资源下第二策略模型的训练吞吐、第一策略模型的推理吞吐、参照模型以及奖励模型的推理吞吐，以最大化总训练吞吐以及推理吞吐为目标进行优化求解，获取推理模块

内各模型的计算资源分配比例，以及，推理模块与训练模块间的计算资源分配比例。

在第一方面的一种实现方式中：

推理模块还用于为训练样本添加时间戳；

经验样本池还用于，检测训练样本的时间戳，当训练样本的时间戳超出预设阈值时，删除训练样本。

根据第一方面的上述实现方式，为训练样本添加时间戳，经验样本池根据训练样本的时间戳自动淘汰过期的训练样本，从而限制生成样本的模型与训练样本模型间的版本差异，保障大规模分布式集群训练下的模型训练收敛能力，提升对齐效果。

在第一方面的一种实现方式中：

推理模块还用于为训练样本添加时间戳；

训练模块还用于，从经验样本池中持续采样训练样本，检验训练样本与当前训练策略模型版本间的差异，当差异超出预设阈值时抛弃训练样本。

根据第一方面的上述实现方式，为训练样本添加时间戳，训练模块获取训练样本时自动比对训练样本与当前模型的版本时间戳，从而限制生成样本的模型与训练样本模型间的版本差异，保障大规模分布式集群训练下的模型训练收敛能力，提升对齐效果。

在第一方面的一种实现方式中，推理模块部署在同构或异构的计算资源上。

在第一方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型，其中：

参照模型以及奖励模型部署在第一类计算资源上，第一策略模型部署在第二类计算资源上。

在第一方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型，其中：

参照模型部署在第一类计算资源上，奖励模型部署在第二类计算资源上，第一策略模型部署在第三类计算资源上。

根据第一方面的上述实现方式，推理模块采用异构式框架，可以有效利用不同类型的计算资源，从而避免云上资源浪费。

在第一方面的一种实现方式中，训练模块部署在同构的计算资源上。

第二方面，本申请提供一种模型训练方法，方法应用于模型系统，模型系统包括推理模块、经验样本池以及训练模块，方法包括：

使用推理模块生成模型训练使用的训练样本；

将推理模块生成的训练样本保存到经验样本池；

使用训练模块，根据经验样本池中保存的训练样本，进行模型的训练，其中，推理模块与训练模块异步并行运行。

在第二方面的一种实现方式中，模型系统用于实现 RLHF 框架。

在第二方面的一种实现方式中，方法还包括：

根据推理模块以及训练模块的吞吐，调整推理模块内各组件的计算资源分配，和/或，调整推理模块以及训练模块的计算资源分配。

在第二方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型，训练模块包含第二策略模型；

根据推理模块以及训练模块的吞吐，调整推理模块内各组件的计算资源分配，和/或，调整推理模块以及训练模块的计算资源分配，包括：

根据单位资源下第二策略模型的训练吞吐、第一策略模型的推理吞吐、参照模型以及奖励模型的推理吞吐，以最大化总训练吞吐以及推理吞吐为目标进行优化求解，获取推理模块内各模型的计算资源分配比例，以及，推理模块与训练模块间的计算资源分配比例。

在第二方面的一种实现方式中，方法还包括：

部署推理模块，其中，推理模块包含并行的多个推理节点；

和/或，

部署训练模块，其中，训练模块包含并行的多个训练节点。

在第二方面的一种实现方式中，方法还包括：

部署推理模块，其中，推理模块部署在同构或异构的计算资源上。

在第二方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型；

部署推理模块，包括：

在第一类计算资源上部署参照模型以及奖励模型，在第二类计算资源上部署第一策略模型。

在第二方面的一种实现方式中，推理模块包含参照模型、奖励模型以及第一策略模型；

部署推理模块，包括：

在第一类计算资源上部署参照模型，在第二类计算资源上部署奖励模型，在第三类计算资源上部署第一策略模型。

在第二方面的一种实现方式中，方法还包括：

部署训练模块，其中，训练模块部署在同构的计算资源上。

在第二方面的一种实现方式中，方法还包括：

在将训练样本保存到经验样本池中之前，为训练样本添加时间戳；

检测经验样本池中训练样本的时间戳，当训练样本的时间戳超出预设阈值时，从经验样本池中删除训练样本。

在第二方面的一种实现方式中，方法还包括：

在将训练样本保存到经验样本池中之前，为训练样本添加时间戳；

从经验样本池中持续采样训练样本，进行模型的训练微调，其中，检验训练样本与当前训练策略模型版本间的差异，当差异超出预设阈值时抛弃训练样本。

第三方面，本申请提供一种电子设备，所述电子设备包括用于存储计算机程序指令的存储器和用于执行计算机程序指令的处理器，其中，当所述计算机程序指令被所述处理器执行时，触发所述电子设备执行如第二方面所述的方法步骤。

第四方面，本申请提供一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机执行第二方面所述的方法。

附图说明

图 1 所示为根据本申请一实施例的模型系统结构示意图；

图 2 所示为根据本申请一实施例的创建模型系统的流程示意图；

图 3 所示为根据本申请一实施例的模型系统结构示意图；

图 4 所示为根据本申请一实施例的 RLHF 方法流程示意图；

图 5 为根据本申请一实施例的电子设备结构示意图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合本申请具体实施例及相应的附图对本申请技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请的实施方式部分使用的术语仅用于对本申请的具体实施例进行解释，而非旨在限定本申请。

大型语言模型（Large Language Model，LLM）的主要目标是基于大量语料数据训练出一个以人为本（诚实、有用、无害）的智能助手。因此，LLM 与人类偏好保持一致具有至关重要的意义。

基于上述原则，针对 LLM 的训练，一种可行的解决方案是采用基于人类反馈的强化学习微调方法

（Reinforcement Learning from Human Feedback，RLHF）。

强化学习是机器学习的范式和方法论之一，用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。RLHF，基于人类反馈的强化学习，是一种将大语言模型与人类偏好相匹配的技术。基于人类偏好数据训练奖励模型建模人类对模型回复的倾向性，通过强化学习算法优化语言模型，以最大化累计奖励，进而实现语言模型输出与人类偏好对齐。

RLHF 利用奖励模型从人类偏好数据中学习价值导向，进而引导语言模型微调过程，是实现人类偏好对齐的关键技术范式。RLHF 能够充分利用人类反馈信息，抽象人类主观偏好，自动优化模型生成更高质量、更符合人类偏好与主流价值观的文本，极大地提高了大语言模型的训练效率和对齐效果。

RLHF 在大语言模型对齐方面具有显著优势，但也引入了许多挑战。首先是采样效率低下，RLHF 采样流程涉及策略模型、价值模型、参考模型和奖励模型多个大模型的训练推理，彼此之间存在串行逻辑，造成设备等待。其次是训练效率低下，强化学习频繁训推切换的特性导致训练样本量受采样效率约束，且多个大模型对显存资源的侵占限制了训练吞吐。

为解决 RLHF 存在的上述问题，一种可行的解决方案是，从显存优化、训推高效切换等多个方向对 RLHF 框架进行优化。

例如，采用 trlX 框架，trlX 框架是一种训练与推理过程串行迭代方案。trlX 框架是围绕抱脸（HuggingFace）生态的分布式 RLHF 框架。trlX 框架的训练过程是基于 HuggingFace 的加速（Accelerate）后端，最大支持 20B 模型的分布式训练。

又例如，采用 ColossalChat 框架，ColossalChat 框架是一种训练与推理过程串行迭代方案。ColossalChat 框架采用低秩自适应（Low-Rank Adaptation，LoRA）与零冗余优化（Zero Redundancy Optimizer，ZeRO）+双子座（Gemini）技术提高内存使用效率、减少通信频率，并能将优化器状态从图形处理器（graphics processing unit，GPU）内存卸载到中央处理器（Central Processing Unit，CPU）内存或硬盘空间，能够以更少的计算资源训练更大的模型。

又例如，采用 DeepSpeedChat 框架，DeepSpeedChat 框架是一种训练与推理过程串行迭代方案。DeepSpeedChat 框架提出一种混合训推引擎实现推理与训练模式间的无缝转换，从而能够利用 DeepSpeed 训练（如 ZeRO、LoRA 等）与推理（如张量并行、高性能算子等）的各种优化手段。

然而，上述优化方案存在以下问题：

1），可扩展性差：上述框架没有充分考虑分布式训练等机制，难以有效对多个模型尺寸与计算资源进行扩展；

2），训推串行迭代：上述框架的训练和推理过程高度耦合，限制了推理优化技术的使用；

3），资源利用率低：RLHF 训推吞吐存在较大差异，上述框架没有考虑训练资源与推理资源的最优分配与异构组合。

针对上述问题，本申请一实施例提供一种模型系统，该模型系统用于实现基于人类反馈的强化学习微调（Reinforcement Learning from Human Feedback，RLHF）方法（RLHF 框架）。

具体的，在本申请一实施例中，模型系统采用异步分布式 RLHF。

异步分布式是指通过异步通信和协调，实现多个节点之间的高效并行处理。异步分布式可以提高系统的吞吐量和可扩展性，同时减少因为通信瓶颈或同步延迟而导致的性能下降。

在本申请一实施例中，模型系统采用异步分布式的 RLHF 框架，即模型系统中，RLHF 框架的训练节点与推理节点独立地并行运行，不进行同步通信与等待。

具体的，在一实施例中，模型系统中 RLHF 框架包含多个并行的训练节点以及多个并行的推理节点，训练节点与推理节点间独立地并行运行，不进行同步通信与等待。

图 1 所示为根据本申请一实施例的模型系统结构示意图。

如图 1 所示，在一实施例中，模型系统的 RLHF 框架包含推理模块 110、训练模块 120（包含多个并行的训练节点）以及经验样本池（Experience Buffer）140。

推理模块 110 用于实现 RLHF 的自采样推理过程。推理模块 110 包含并行的推理节点，其负责自

采样生成模型训练使用的样本数据。

推理模块 110 包括参照模型（Reference Model，Ref Model）111、奖励模型（Reward Model，RM）112、策略模型（Policy Model）113。

参照模型 111，在 RLHF 中指用于初始化策略模型的原始语言模型，其用于针对给定的用户提问，提供一个参考答案，与策略模型 113 生成的模型回复进行对比，限制策略模型 113 的变化幅度。

奖励模型 112，RLHF 是奖励引导下的策略优化过程，奖励模型 112 以{用户提问，模型输出}为输入，输出标量分值，用于评估策略模型 113 生成回复的好坏。

策略模型 113，为 RLHF 中待优化的大语言模型，以用户提问为输入，输出模型回复。RLHF 的最终目的为优化策略模型使其输出高质量文本并与人类偏好对齐。

在推理模块 110 的推理阶段，从数据集采样用户问题输入策略模型 113，前向推理得到问答对，将问答对传输给参照模型 111 与奖励模型 112 分别推理计算 KL 惩罚与奖励值，得到训练样本，并存储到经验样本池 140。

训练模块 120 用于 RLHF 微调训练过程。训练模块 120 包含多个并行的训练节点，用于基于强化学习方法进行模型的训练微调。

在训练阶段，训练模块 120 加载策略模型，从经验样本池 140 中持续进行采样，并基于强化学习算法（如近端策略优化算法（Proximal Policy Optimization，PPO））进行模型训练微调。

PPO 是一种基于演员兼评论家（Actor-Critic）的强化学习算法，用于训练策略模型完成复杂任务。

经验样本池 140 存储在计算资源内存中，用于接收推理模块 110 构造的样本数据，并为训练模块 120 提供 RLHF 微调所需的训练数据。

本申请一实施例提供的异步分布式 RLHF 框架采用训练模块 120 与推理模块 110 分离（训推分离）的异步异构多节点集群架构，具备良好的可扩展性，有效解决大模型 RLHF 方法采样效率低下的问题，提升模型训练的训练效率。

根据本申请一实施例，RLHF 框架采用异步分布式框架，使得 RLHF 训练与推理过程解耦，采用训练节点与推理节点异步并行的运行方式，具备良好的可扩展性，同时能够针对训练与推理分别灵活应用现有的优化技术。

可选的，在一实施例中，模型系统的 RLHF 框架采用异构分布式框架。异构分布式框架是指由不同类型或者不同平台的节点所组成的框架，异构分布式框架同时利用不同类型或者不同平台的节点各自的优势进行高效的并行计算。

具体的，在一实施例中，推理模块 110 兼容同构或异构的计算资源配置。例如，RLHF 框架的推理节点计算资源可同时选用 GPU 与神经网络处理器（Neural-network Processing Unit，NPU）进行并行计算。

在一实施例中，在推理模块 110 中，策略模型 113 独立部署在计算资源上，参照模型 111 与奖励模型 112 共同部署在同一计算资源上。

根据本申请一实施例，RLHF 框架采用异构分布式框架，可以有效利用不同类型的计算资源（如 GPU、NPU、CPU），从而避免云上资源浪费。

进一步的，在一实施例中，模型系统还包含协调控制器（Controller）130。协调控制器（Controller）130 用于调整模型系统中 RLHF 框架各组件的资源占用。

具体的，协调控制器 130 根据训练模块 120 与推理模块 110 的吞吐，自适应动态调整推理模块 110 内各组件的计算资源分配（例如，调整参照模型 111、奖励模型 112、策略模型 113 的计算资源分配比例），以及，调整训练模块 120 与推理模块 110 的计算资源分配，实现数据流的动态平衡，确保训练模块 120 与推理模块 110 满负荷运行，实现计算资源利用率的最大化。

本申请一实施例提供的异步分布式 RLHF 框架设计了自适应资源协调机制，提供动态调整推理节点内计算资源比例与训推节点间计算资源比例功能，有效提升 RLHF 训练吞吐与资源利用率。

图 2 所示为根据本申请一实施例的创建模型系统的流程示意图。

在一实施例中，电子设备执行如图 2 所示的下述流程以构建并训练图 1 所示的 RLHF 框架。

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N^{infer}。

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N^{train}。

S230，构建协调控制器 130，使用协调控制器 130 计算推理模块 110 与训练模块 120 的资源配比的优化。

具体的，协调控制器 130 根据单位资源下策略模型训练吞吐t_p^{train}、策略模型推理吞吐t_p^{infer}、参照/奖励模型推理吞吐t_r^{infer}，以最大化总训练吞吐以及推理吞吐为目标进行优化求解，得到推理模块 110 内最优的策略模型与参照/奖励模型间计算资源配比，以及训练模块 120 与推理模块 110 间最优的计算资源配比。

在本申请实施例的描述中，训练吞吐是指单位时间内可以处理的最大样本数量或词元数量。推理吞吐是指单位时间内可以推理的最大样本数量或词元数量。

优选地，在一实施例中，协调控制器 130 进行计算资源配比的优化求解，优化目标为在给定训推资源下最大化 RLHF 框架的训推吞吐。

具体地，建立优化函数如下：

$$\max_{N_p^{infer},\ N_r^{infer},\ N_p^{train}} \min(T^{train}, T^{infer})。（式 1）$$

式 1 中，N_p^{train}代表训练资源数量，N_p^{infer}代表策略模型推理资源数量、N_r^{infer}代表参照/奖励模型推理资源数量。T^{train}与T^{infer}分别是训练模块与推理模块的总吞吐量。

训练与推理模块间，为避免训练模块等待推理模块导致的资源闲置问题，要求训练总吞吐量不大于推理总吞吐量，具体约束如下：

$$\begin{cases} T^{train} \le T^{infer} \\ T^{train} = t_p^{train} \cdot N_p^{train} \cdot K_t \\ T^{infer} = \min(T_p^{infer}, T_r^{infer}) \\ N_p^{train} \le N^{train} \end{cases}。（式 2）$$

式 2 中，t_p^{infer}与N_p^{infer}分别为策略模型的单卡推理吞吐与卡数，t_r^{infer}与N_r^{infer}分别为参照/奖励模型的单卡推理吞吐与卡数，K_i为模型推理所需的最小计算卡数。

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。

具体的，在网络状态发生较大波动时，协调控制器 130 可以对推理模块 110 以及训练模块 120 的资源配比与数量进行动态优化。

根据本申请一实施例，RLHF 框架采用自适应资源协调，动态调整推理节点内的计算资源比例，最优化策略模型与参照/奖励模型吞吐；动态调整训练与推理节点间的计算资源配比，避免训练资源等待与推理数据浪费。

可选的，在一实施例中，推理模块 110 的推理阶段生成的训练样本添加时间戳。

添加时间戳后的训练样本被存入经验样本池 140。样本数据格式：

$$x_s = \{\text{prompt}, \text{response}, \text{logprob}, \text{value}, \text{reward}, \text{timestamp}\}。\quad（式 3）$$

在一实施例中，经验样本池 140 中的样本数据的生成时间在超过阈值 Staleness（例如，10 个训练周期）后，经验样本池 140 会自动删除该数据。

根据本申请一实施例，RLHF 框架采用时间戳，经验样本池根据时间戳自动淘汰过期的训练样本，从而限制生成样本的模型与训练样本模型间的版本差异，保障大规模分布式集群训练下的 RLHF 训练收敛能力，提升对齐效果。

在另一实施例中，训练模块 120 从经验样本池 140 中持续采样训练样本，检验训练样本的时间戳（timestamp）与当前所训练策略模型版本间的差异程度，若超出给定阈值 Staleness（例如，10 个版本迭代周期）则直接抛弃该数据。

根据本申请一实施例，RLHF 框架采用时间戳，训练模块获取训练样本时自动比对训练样本与当前模型的版本时间戳，限制生成样本的模型与训练样本模型间的版本差异，保障大规模分布式集群训练下的 RLHF 训练收敛能力，提升对齐效果。

在另一实施例中，经验样本池 140 中的样本数据的生成时间在超过阈值 Staleness（例如，10 个训练周期）后，经验样本池 140 会自动删除该数据。并且，训练模块 120 从经验样本池 140 中持续采样训练样本，检验训练样本的时间戳（timestamp）与当前所训练策略模型版本间的差异程度，若超出给定阈值 Staleness（例如，10 个版本迭代周期）则直接抛弃该数据。

本申请实施例所提出的 RLHF 框架可应用于云上单计算资源（仅 GPU 或 NPU）与混合计算资源（GPU+NPU）或（GPU+NPU+CPU）下的 RLHF 训练场景。下面结合百亿参数量大小的 LLM 实施例对本发明进行详细说明。以下实施例将有助于本领域的技术人员进一步理解本发明，但不以任何形式限制本发明。应当指出的是，对本领域的普通技术人员来说，在不脱离本发明构思的前提下，还可以做出若干变形和改进。这些都属于本发明的保护范围。

图 3 所示为根据本申请一实施例的模型系统结构示意图。

如图 3 所示，在一实施例中，模型系统由推理模块 310、训练模块 320、经验样本池 340 以及协调控制器 330 组成。

经验样本池 340 存储在计算资源内存中，其用于存储强化学习经验样本。

在推理过程中，推理模块 310 自采样生成训练数据，存储在经验样本池 340 中。

推理模块 310 经验样本池 340 接收推理模块 310 构造的样本数据，并为训练模块 320 提供 RLHF 微调所需的训练数据。

如图 3 所示，推理模块 310 包含策略模型 313、参照模型 311 以及奖励模型 312。

推理模块 310 的推理过程为，策略模型 313 根据原始样本池 314 的激发（prompt），自回归生成响应（responses）。激发（prompt）以及响应（responses）被传输给参照模型 311 与奖励模型 312，以分别计算 KL 惩罚与奖励值，并将生成的样本{prompt, response, logprob, value, reward, timestamp}存储在经验样本池 340 中。

训练模块 320 包含策略模型 321 以及设置模型（SET Model）322。

训练模块 320 使用有监督微调模型（supervised fine-tunning Model，SFT Model）322 对策略模型 321 进行模型初始化。

训练模块 320 从经验样本池 340 中持续采样训练样本，检验样本 timestamp 与当前所训练策略模型版本间的差异程度，若超出给定阈值 Staleness=2 则直接抛弃该数据。

进一步的，分布式并行 RLHF 框架中训练模块 320 与推理模块 310 的策略模型（策略模型 313 与策略模型 321）会存在一定差异，经验样本池 340 存储样本中的 timestamp 为当前推理模块 310 中策略模型 313 的时间戳，用于记录采样样本来源。

在单计算资源场景下，训练模块 320 与推理模块 310 都使用同一种硬件资源，如 GPU 或 NPU。

在混合计算资源场景下，训练模块 320 使用单一种类计算资源，如 GPU 或 NPU。推理模块 310 可使用多种类别计算资源，如 GPU+NPU+CPU。

具体的，在构建 RLHF 框架时， 选择一种或多种计算资源用于推理模块 310。基于并行训练的通信问题，选择一种计算资源用于训练模块 320。

例如，考虑到百亿 LLM 推理显存占用与计算量需求，华为云平台中可选的云上计算资源包括

NVIDIA A800 GPU 资源与 Ascend D910 NPU 计算资源，两者在百亿模型场景下的最小推理资源都为单卡。Ascend D910 NPU 是华为自研 AI 计算芯片。NVIDIA A800 GPU 是英伟达公司研发的图形处理器。可选择一种或多种计算资源用于推理模块。由于并行训练的通信问题，仅可选择一种计算资源用于训练模块。

可选的，在一实施例中，策略模型 313 独占计算资源，奖励模型 312 与参照模型 311 共享计算资源。在另一实施例中，奖励模型 312、参照模型 311 独占计算资源。

协调控制器 330 部署在 CPU 上。协调控制器 330 根据训练模块 320 与推理模块 310 吞吐，自适应动态调整推理模块 310 内计算资源比例，以及训练模块 320 与推理模块 310 间计算资源比例（调整推理和训练节点的比率），实现数据流的动态平衡，确保训练模块 320 与推理模块 310 满负荷运行，实现计算资源利用率的最大化。

具体的，在一实施例中，对于推理模块 310 内部，要求策略模型 313 总吞吐量小于等于参照模型 311/奖励模型 312 总吞吐量，参照式 2，即：

$$\begin{cases} T_p^{infer} \le T_r^{infer} \\ T_p^{infer} = t_p^{infer} \cdot N_p^{infer} \cdot 1 \\ T_r^{infer} = t_r^{infer} \cdot N_r^{infer} \cdot 1 \\ N_p^{infer} + N_r^{infer} \le N^{infer} \end{cases}。 \quad （式 4）$$

式 4 中，t_p^{infer}与N_p^{infer}分别为策略模型的单卡推理吞吐与卡数，t_r^{infer}与N_r^{infer}分别为参照/奖励模型的单卡推理吞吐与卡数，T_p^{infer}与T_r^{infer}分别为策略模型与参照/奖励模型的推理总吞吐量。

对于训练模块 320 与推理模块 310 间，要求训练总吞吐量不大于推理总吞吐量，即：

$$\begin{cases} T^{train} \le T^{infer} \\ T^{train} = t_p^{train} \cdot N_p^{train} \cdot 8 \\ T^{infer} = \min\left(T_p^{infer}, T_r^{infer}\right) \\ N_p^{train} \le N^{train} \end{cases}。 \quad （式 5）$$

式 5 中，T^{train}与T^{infer}分别表示训练模块与推理模块的总吞吐量，t_p^{train}与N_p^{train}分别为模型训练的单卡吞吐与卡数。

协调控制器 330 优化目标为在给定训推资源下最大化 RLHF 框架的训推吞吐，计算求解出最优的策略模型推理资源数量N_p^{infer}、参照/奖励模型推理资源数量N_r^{infer}与训练资源数量N_p^{train}，即参照式 1。

图 4 所示为根据本申请一实施例的 RLHF 方法流程示意图。

S410，选取所需的推理计算资源种类，设置资源数量上限（参照 S210）。

S420，选取所需的训练计算资源种类，设置资源数量上限（参照 S220）。

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（参照 S230）。

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。

具体的，在 S440 中，最初使用 S430 计算的初始资源配比。

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（参照 S230）。

在 S450 之后返回 S440，在 S440 中使用 S450 计算的最优资源配比。

本申请实施例提供的 RLHF 框架，除应用于大模型 RLHF 场景外，也适用于大模型场景下的通用强化学习训练场景，用以实现强化学习中训练与推理节点资源的灵活扩充，提升强化学习训练效率。

本领域技术人员知道，除了以纯计算机可读程序代码方式实现本发明提供的系统及其各个装置以外，完全可以通过将方法步骤进行逻辑编程来使得本发明提供的系统及其各个装置以逻辑门、开关、专用集成电路、可编程逻辑控制器以及嵌入式微控制器等的形式来实现相同功能。所以，本发明提供

的系统及其各项装置可以被认为是一种硬件部件，而对其内包括的用于实现各种功能的装置也可以视为硬件部件内的结构；也可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件部件内的结构。

此处公开的仅为本发明的优选实施例，本说明书选取并具体描述这些实施例，是为了更好地解释本发明的原理和实际应用，并不是对本发明的限定。任何本领域技术人员在说明书范围内所做的修改和变化，均应落在本发明所保护的范围内。

在本申请实施例的描述中，为了描述的方便，描述着色装置时以功能分为各种模块分别描述，各个模块的划分仅仅是一种逻辑功能的划分，在实施本申请实施例时可以把各模块的功能在同一个或多个软件和/或硬件中实现。

具体的，本申请实施例所提出的装置在实际实现时可以全部或部分集成到一个物理实体（例如，GPU 或其他类型处理器）上，也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现；也可以全部以硬件的形式实现；还可以部分模块以软件通过处理元件调用的形式实现，部分模块通过硬件的形式实现。例如，检测模块可以为单独设立的处理元件，也可以集成在电子设备的某一个芯片中实现。其它模块的实现与之类似。此外这些模块全部或部分可以集成在一起，也可以独立实现。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。

例如，以上这些模块可以是被配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路（Application Specific Integrated Circuit，ASIC），或，一个或多个数字信号处理器（Digital Singnal Processor, DSP），或，一个或者多个现场可编程门阵列（Field Programmable Gate Array, FPGA）等。再如，这些模块可以集成在一起，以片上装置（System-On-a-Chip，SOC）的形式实现。

本申请一实施例还提出了一种电子设备。该电子设备用于执行本申请实施例所述的方法流程或部分的方法流程。

图 5 为根据本申请一实施例的电子设备结构示意图。

如图 5 所示，电子设备 2500 包括用于存储计算机程序指令的存储器 2502 和用于执行程序指令的处理器 2501，其中，当该计算机程序指令被该处理器 2501 执行时，触发电子设备 2500 执行如本申请实施例所述的方法步骤。

具体的，在本申请一实施例中，上述一个或多个计算机程序被存储在上述存储器 2502 中，上述一个或多个计算机程序包括指令，当上述指令被上述电子设备 2500 执行时，使得上述电子设备 2500 执行本申请实施例所述的方法步骤。

可以理解的是，本申请实施例对电子设备 2500 的结构描述并不构成对电子设备 2500 的具体限定。在本申请另一些实施例中，电子设备 2500 可以包括处理器 2501 以及存储器 2502 以外的其他部件。

处理器 2501 可以是片上装置 SOC，该处理器 2501 中可以包括中央处理器（Central Processing Unit，CPU），还可以进一步包括其他类型的处理器。

处理器2501涉及的处理器可以例如包括CPU、DSP、微控制器或数字信号处理器，还可包括GPU、嵌入式神经网络处理器（Neural-network Process Units，NPU）和图像信号处理器（Image Signal Processing，ISP），处理器还可包括必要的硬件加速器或逻辑处理硬件电路，如 ASIC，或一个或多个用于控制本申请技术方案程序执行的集成电路等。此外，处理器可以具有操作一个或多个软件程序的功能，软件程序可以存储在存储介质中。

处理器 2501 可以包括一个或多个处理单元。例如：处理器可以包括应用处理器（application processor，AP），调制解调处理器，图形处理器（graphics processing unit，GPU），图像信号处理器（image signal processor，ISP），控制器，视频编解码器，数字信号处理器（digital signal processor，DSP），基带处理器，和/或神经网络处理器（neural-network processing unit，NPU）等。其中，不同的处理单元可以是独立的部件，也可以集成在一个或多个处理器中。在一些实施例中，电子设备2500也可以包括一个或多个处理器 2501。其中，控制器可以根据指令操作码和时序信号，产生操作控制信号，完成取指令和执行指令的控制。

在一些实施例中，处理器 2501 可以包括一个或多个接口。接口可以包括集成电路间（inter-

integrated circuit，I2C）接口，集成电路间音频（integrated circuit sound，I2S）接口，脉冲编码调制（pulse code modulation，PCM）接口，通用异步收发传输器（universal asynchronous receiver/transmitter，UART）接口，移动产业处理器接口（mobile industry processor interface，MIPI），通用输入输出（general-purpose input/output，GPIO）接口，SIM 卡接口，和/或 USB 接口等。其中，USB 接口是符合 USB 标准规范的接口，具体可以是 Mini USB 接口，Micro USB 接口，USB Type C 接口等。USB 接口可以用于连接充电器为电子设备充电，也可以用于电子设备与外围设备之间传输数据。

电子设备 2500 还可以包括外部存储器接口，外部存储器接口用于连接外部存储卡，例如 Micro SD 卡，实现扩展电子设备的存储能力。外部存储卡通过外部存储器接口与处理器 2501 通信，实现数据存储功能。例如将音乐，视频等文件保存在外部存储卡中。

存储器 2502 可以包括代码存储区和数据存储区。其中，代码存储区可存储操作系统。数据存储区可存储电子设备 2500 使用过程中所创建的数据等。此外，存储器 2502 可以包括高速随机存取存储器，还可以包括非易失性存储器，例如一个或多个磁盘存储部件，闪存部件，通用闪存存储器（universal flash storage，UFS）等。

存储器 2502 可以是只读存储器（read-only memory，ROM）、可存储静态信息和指令的其它类型的静态存储设备、随机存取存储器（random access memory，RAM)或可存储信息和指令的其它类型的动态存储设备，也可以是电可擦可编程只读存储器（electrically erasable programmable read-only memory，EEPROM）、只读光盘（compact disc read-only memory，CD-ROM）或其他光盘存储、光碟存储（包括压缩光碟、激光碟、光碟、数字通用光碟、蓝光光碟等）、磁盘存储介质或者其它磁存储设备，或者还可以是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何计算机可读介质。

处理器 2501 和存储器 2502 可以合成一个处理装置，更常见的是彼此独立的部件。

本申请一实施例还提出了一种电子芯片。该电子芯片用于执行本申请实施例所述的方法流程或部分的方法流程。例如，该电子芯片可以为 GPU、CPU 或 NPU。

具体的，电子芯片包括用于执行程序指令的处理器，当计算机程序指令被该处理器执行时，触发电子芯片执行如本申请实施例所述的方法步骤。电子芯片的处理器可以参照上述电子设备的处理器。

可选的，本申请实施例阐明的设备、装置、模块，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。

本领域内的技术人员应明白，本申请实施例可提供为方法、装置、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质上实施的计算机程序产品的形式。

在本申请所提供的几个实施例中，任一功能如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。

具体的，本申请一实施例中还提供一种计算机可读存储介质，该计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机执行本申请实施例提供的方法。

本申请一实施例还提供一种计算机程序产品，该计算机程序产品包括计算机程序，当其在计算机上运行时，使得计算机执行本申请实施例提供的方法。

本申请中的实施例描述是参照根据本申请实施例的方法、设备（装置）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计

算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

还需要说明的是，本申请实施例中，“至少一个”是指一个或者多个，“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示单独存在A、同时存在A和B、单独存在B的情况。其中A，B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项”及其类似表达，是指的这些项中的任意组合，包括单项或复数项的任意组合。例如，a，b和c中的至少一项可以表示：a，b，c，a和b，a和c，b和c或a和b和c，其中a，b，c可以是单个，也可以是多个。

本申请实施例中，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本申请可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本申请，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

本申请中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

本领域普通技术人员可以意识到，本申请实施例中描述的各单元及算法步骤，能够以电子硬件、计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置、装置和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

以上所述，仅为本申请的具体实施方式，任何熟悉本技术领域的技术人员在本申请公开的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。本申请的保护范围应以所述权利要求的保护范围为准。

权 利 要 求 书

1.一种模型训练系统，其特征在于，所述模型训练系统包括：

推理模块，其用于生成模型训练使用的训练样本；

经验样本池，其用于保存所述训练样本；

训练模块，其用于根据所述经验样本池中保存的所述训练样本，进行模型的训练，其中，所述推理模块与所述训练模块异步并行运行。

2.根据权利要求 1 所述的模型训练系统，其特征在于，所述模型训练系统还包括：

协调控制器，其用于根据所述推理模块以及所述训练模块的吞吐，调整所述推理模块内各组件的计算资源分配，和/或，调整所述推理模块以及所述训练模块的计算资源分配。

3.根据权利要求 2 所述的模型训练系统，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型，所述训练模块包含第二策略模型；

所述协调控制器用于，根据单位资源下所述第二策略模型的训练吞吐、所述第一策略模型的推理吞吐、所述参照模型以及所述奖励模型的推理吞吐，以最大化总训练吞吐以及推理吞吐为目标进行优化求解，获取所述推理模块内各模型的计算资源分配比例，以及，所述推理模块与所述训练模块间的计算资源分配比例。

4.根据权利要求 1 所述的模型训练系统，其特征在于：

所述推理模块还用于为所述训练样本添加时间戳；

所述经验样本池还用于，检测所述训练样本的时间戳，当所述训练样本的时间戳超出预设阈值时，删除所述训练样本。

5.根据权利要求 1 所述的模型训练系统，其特征在于：

所述推理模块还用于为所述训练样本添加时间戳；

所述训练模块还用于，从所述经验样本池中持续采样所述训练样本，检验所述训练样本与当前训练策略模型版本间的差异，当所述差异超出预设阈值时抛弃所述训练样本。

6.根据权利要求 1-5 中任一项所述的模型训练系统，其特征在于，所述推理模块部署在同构或异构的计算资源上。

7.根据权利要求 6 所述的模型训练系统，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型，其中：

所述参照模型以及所述奖励模型部署在第一类计算资源上，所述第一策略模型部署在第二类计算资源上。

8.根据权利要求 7 所述的模型训练系统，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型，其中：

所述参照模型部署在第一类计算资源上，所述奖励模型部署在第二类计算资源上，所述第一策略模型部署在第三类计算资源上。

9.根据权利要求 1-5 中任一项所述的模型训练系统，其特征在于，所述训练模块部署在同构的计算资源上。

10.一种模型训练方法，其特征在于，所述方法应用于模型系统，所述模型系统包括推理模块、经验样本池以及训练模块，所述方法包括：

使用所述推理模块生成模型训练使用的训练样本；

将所述推理模块生成的所述训练样本保存到所述经验样本池；

使用所述训练模块，根据所述经验样本池中保存的所述训练样本，进行模型的训练，其中，所述推

理模块与所述训练模块异步并行运行。

11.根据权利要求 10 所述的方法，其特征在于，方法还包括：

根据所述推理模块以及所述训练模块的吞吐,调整所述推理模块内各组件的计算资源分配,和/或，调整所述推理模块以及所述训练模块的计算资源分配。

12.根据权利要求 11 所述的方法，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型，所述训练模块包含第二策略模型；

所述根据所述推理模块以及所述训练模块的吞吐，调整所述推理模块内各组件的计算资源分配，和/或，调整所述推理模块以及所述训练模块的计算资源分配，包括：

根据单位资源下所述第二策略模型的训练吞吐、所述第一策略模型的推理吞吐、所述参照模型以及所述奖励模型的推理吞吐，以最大化总训练吞吐以及推理吞吐为目标进行优化求解，获取所述推理模块内各模型的计算资源分配比例，以及，所述推理模块与所述训练模块间的计算资源分配比例。

13.根据权利要求 10 所述的方法，其特征在于，所述方法还包括：

在将所述训练样本保存到所述经验样本池中之前，为所述训练样本添加时间戳；

检测所述经验样本池中训练样本的时间戳，当所述训练样本的时间戳超出预设阈值时，从所述经验样本池中删除所述训练样本。

14.根据权利要求 10 所述的方法，其特征在于，所述方法还包括：

在将所述训练样本保存到所述经验样本池中之前，为所述训练样本添加时间戳；

从所述经验样本池中持续采样所述训练样本，进行模型的训练微调，其中，检验所述训练样本与当前训练策略模型版本间的差异，当所述差异超出预设阈值时抛弃所述训练样本。

15.根据权利要求 10-14 中任一项所述的方法，其特征在于，所述方法还包括：

部署所述推理模块，其中，所述推理模块部署在同构或异构的计算资源上。

16.根据权利要求 15 所述的方法，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型；

所述部署所述推理模块，包括：

在第一类计算资源上部署所述参照模型以及所述奖励模型，在第二类计算资源上部署所述第一策略模型。

17.根据权利要求 15 所述的方法，其特征在于，所述推理模块包含参照模型、奖励模型以及第一策略模型；

所述部署所述推理模块，包括：

在第一类计算资源上部署所述参照模型，在第二类计算资源上部署所述奖励模型，在第三类计算资源上部署所述第一策略模型。

18.根据权利要求 10-14 中任一项所述的方法，其特征在于，所述方法还包括：

部署所述训练模块，其中，所述训练模块部署在同构的计算资源上。

19.一种电子设备，其特征在于，所述电子设备包括用于存储计算机程序指令的存储器和用于执行计算机程序指令的处理器，其中，当所述计算机程序指令被所述处理器执行时，触发所述电子设备执行如权利要求 10-18 中任一项所述的方法步骤。

20.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序，当其在计算机上运行时，使得计算机执行如权利要求 10-18 中任一项所述的方法步骤。