大语言模型的管理系统

本申请要求在2025年03月13日提交中国专利局、申请号为202510295398.4的中国专利申请的优先权，以上申请的全部内容通过引用结合在本申请中。

技术领域

本申请涉及人工智能技术领域，例如涉及一种大语言模型的管理系统。

背景技术

由于汽车金融业务涉及多种场景和数据源，为保证数据金融数据的安全，推理服务往往需要跨至少一个集群和环境进行协同管理。

然而，目前的协同管理技术手段在多集群间的通信与资源调度方面存在较大的复杂性，难以实现高效的资源利用和推理性能保障。同时，多环境的隔离与部署策略缺乏统一的管理方案，容易导致环境间资源冲突和推理结果不可靠。

发明内容

本申请提供了一种大语言模型的管理系统，所述大语言模型的管理系统包括配置模块、至少一个集群以及日志查询模块：所述配置模块与至少一个集群通过数据操作服务进行通信连接，所述至少一个集群与所述日志查询模块通过网络通信连接；

所述配置模块设置为响应于用户触发操作，基于所述用户触发操作得到配置信息，并将所述配置信息发送到至少一个集群中；

针对每个集群，所述集群设置为接收所述配置信息，并对所述配置信息进行配置处理后得到与所述配置信息对应的大语言模型；其中，所述配置信息包括大语言模型的部署信息以及模型信息；

所述日志查询模块设置为接收查询请求，基于所述查询请求查询所述集群得到所述大语言模型的调用记录信息。

可选的，所述配置模块包括模型仓库单元和应用管理单元；

所述模型仓库单元设置为响应所述用户触发操作提供多种功能类型的大语言模型；

所述应用管理单元设置为响应所述用户触发操作选择对应的部署信息。

可选的，所述集群包括目标环境库；

所述配置模块设置为将所述大语言模型部署至所述大语言模型对应的集群以及目标环境库中。

可选的，所述目标环境库包括如下至少之一：生产环境库或测试环境库。

可选的，所述配置模块设置为将大语言模型发布至测试环境库进行测试验证，并将测试验证通过的大语言模型部署至生产环境库进行生产使用。

可选的，所述集群包括引擎服务单元和模型推理服务单元；

所述引擎服务单元设置为将发布至目标环境库的大语言模型应用信息按照设定时长加载到所述模型推理服务单元中；

所述模型推理服务单元设置为根据所述大语言模型应用信息得到大语言模型的服务信息，并将大语言模型的服务信息进行注册操作，得到大语言模型的注册信息。

可选的，所述集群包括应用调度服务单元；

所述引擎服务单元设置为将所述大语言模型应用信息按照设定时长加载到所述应用调度服务单元中；

所述应用调度服务单元设置为解析所述大语言模型的应用信息，以根据所述应用信息和所述注册信息调用所述模型推理服务单元。

可选的，所述模型推理服务单元还设置为提供大语言模型应用推理服务的接口，以使得用户通过接口调用所述大语言模型进行使用。

可选的，所述集群还包括日志推送单元；

所述日志推送单元设置为将所述模型推理服务单元和所述应用调度服务单元的执行记录推送至消息队列中。

可选的，所述集群还包括日志解析单元；

所述日志解析单元设置为对所述消息队列中的执行记录进行解析并保存至所述集群的日志库中，以供所述日志查询模块查看。

附图说明

图 1 是根据本申请实施例提供的一种大语言模型的管理系统的结构示意图；

图 2 是根据本申请实施例提供的一种大语言模型的管理系统的结构示例图。

具体实施方式

本申请的说明书和权利要求书及上述附图中的术语“目标”、“原始”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”及其任何变形，意图在于覆盖不排他的包含，例如，除了包含本申请实施例示出的一系列步骤或单元的过程、方法、系统、产品或设备，还可包括没有清楚地列出的这一系列步骤或单元的其它过程、方法、系统、产品和设备，或对于这些过程、方法、系统、产品或设备固有的其它步骤或单元。

图 1 是根据本申请实施例提供的一种大语言模型的管理系统的结构示意图，如图 1 所示，所述大语言模型的管理系统包括配置模块 110、至少一个集群 120 以及日志查询模块 130；所述配置模块与至少一个集群通过数据操作服务进行通信连接，所述至少一个集群与所述日志查询模块通过网络通信连接；其中，数据操作服务可以是指基于设定的通信机制或服务，用于负责处理大语言模型与各个集群之间的数据传输和交互。本实施例中数据操作服务可以是指通过设定的通信接口实现配置模块与至少一个集群的通信。示例性的，数据操作服务可以为运维（Operations，op）服务。本实施例中的配置模块可以是通过云端页面配置的方式呈现给用户，以使得用户通过触发操作选取配置内容得到对应的配置信息，如配置大语言模型的部署位置，所需要部署的具体对象以及存储位置等信息。本实施例中的大语言模型可以针对于汽车金融风控场景下的大语言模型。

所述配置模块 110 设置为响应于用户触发操作，基于所述用户触发操作得到配置信息，并将所述配置信息发送到至少一个集群中。

其中，所述配置信息包括大语言模型的部署信息以及模型信息；部署信息可以包括部署对象和具体的部署位置等信息。部署对象可以是部署指定环境以

及集群等对象。部署位置可以是大语言模型最终需要部署的位置。模型信息可以是指大语言模型的存储的位置信息。其中，本实施例中至少一个集群可以包括集群A、集群B、集群C...集群N等等，具体数量不作限制。

本实施例中用户通过触发操作选择大语言模型的部署信息以及模型信息等内容，配置模块响应于该触发操作，根据该触发操作得到用户选择的配置信息，并将配置信息发送到至少一个集群中。本实施例中大语言模型的管理系统采用统一的多集群多环境的管理方式，具体将环境库与集群混合部署，集群中存在环境库，但是又不需要具备完整的环境库，将多集群的隔离与多环境部署策略相结合，解决环境间资源冲突问题。可以理解的，本实施例中的至少一个集群中可以包括一个或者多个指定环境库，以便于在对大语言模型的部署时，选择对应的集群和其包含的指定环境库。其中，环境库可以包括但不限于测试环境库和生产环境库。

本实施例中在基于配置模块对大语言模型的部署时，可以先选择环境库，再选择大语言模型要部署的集群的方式，以确定大语言模型最终部署的集群对象和环境对象；本实施例中可以通过数据操作服务解决部署时的页面配置服务和各集群网络隔离的问题，通过各集群日志分别采集的方式解决各隔离集群调用日志反显的问题。金融风控场景的大语言模型为尽可能保证推理结果的可靠性，需要先发布部署测试环境验证，再将相同的大语言模型服务应用部署到生产环境投入使用。

针对每个集群120，所述集群120设置为接收所述配置信息，并对所述配置信息进行配置处理后得到与所述配置信息对应的大语言模型。

其中，集群可基于所包含的引擎服务对配置信息进行配置处理后得到对应的大语言模型。引擎服务可以是用于处理大语言模型相关任务的软件服务。本实施例中的集群中所包含的引擎服务通常可以包括多种功能服务，用于处理大语言模型的不同方面。示例性的，引擎服务可以包括训练服务、推理服务、优化服务等，每一服务都负责特定的任务。

本实施例中各个集群在接收到配置模块发送的大语言模型对应的配置信息后，可以根据各个集群所包含的环境库不同，先将大语言模型的相关应用信息经过环境库进行对应的处理，再进行引擎服务等其他服务的处理，从而可以得

到配置后的大语言模型。

所述日志查询模块 130 设置为接收查询请求，基于所述查询请求查询所述集群得到所述大语言模型的调用记录信息。

其中，查询请求可以是在每个集群得到对应的大语言模型后，用户发出的对各个集群内的使用情况进行查询的请求。本实施例中每个集群中都具有各自的日志推送功能。每个集群在每次调用集群内的不同服务时，对应生成一个调用记录，然后通过日志推送功能将该调用记录推送至调用明细单元中，以便于日志查询模块进行查询。其中，配置模块中包括调用明细单元，用于将不同集群生成的调用记录展示在云端页面。本实施例中，响应于接收到用户发出的日志查询请求，通过该查询请求查询对应的至少一个集群中所对应的调用记录信息以及其他使用信息等等。

在本实施例中，可选的，所述配置模块包括模型仓库单元和应用管理单元；所述模型仓库单元设置为响应用户触发操作提供多种功能类型的大语言模型；所述应用管理单元设置为响应用户触发操作选择对应的部署信息。

本实施例的大语言模型的管理系统的配置模块包括模型仓库单元和应用管理单元。模型仓库单元里面可以包括有各种不同功能类型的大语言模型，设置为响应于用户的触发操作，将大语言模型上传保存至云端；应用管理单元设置为响应于用户的触发操作选择对应的集群和环境，以将大语言模型应用部署到对应的集群和环境中。

本实施例可以通过云端页面配置模型信息和大语言模型的模型文件等信息，灵活地确定出满足用户需求的模型配置。

在本实施例中，可选的，所述集群包括目标环境库；所述配置模块将所述大语言模型部署至大语言模型对应的集群以及目标环境库中。

其中，所述目标环境库包括生产环境库和/测试环境库。测试环境库可以用于对大语言模型进行应用测试操作。生产环境库可以用于生产大语言模型服务应用。可以理解的，不同的集群中可以包括不同的目标环境库，示例性的，集群 A 可以包括生产环境库；集群 B 可以包括测试环境库；集群 C 可以包括生产环境库和测试环境库。本实施例中可以根据集群的具体定位设置等信息确定其

所包含的目标环境库。

本实施例中可以通过配置模块将大语言模型部署到对应的至少一个集群以及集群所对应的目标环境库中。本实施例中通过这样的设置，可以对多个集群和多个环境进行管理，以便于为用户提供种类不同的部署设置，以满足用户需求。

在本实施例中，可选的，所述配置模块将大语言模型发布至测试环境库进行测试验证，并将测试验证通过的大语言模型部署至生产环境进行生产使用。

本实施例中配置模块可以将大语言模型发送至对应的测试环境库中，进行大语言模型的应用功能的测试验证，并将通过测试验证的大语言模型发布至对应的生产环境库中，以通过生产环境库将大语言模型服务应用于生产使用。

本实施例中通过这样的设置，需要先发布部署测试环境验证，再将相同的大语言模型服务应用部署到生产环境中投入使用，从而保证了金融风控场景的大语言模型推理结果的可靠性。

在本实施例中，可选的，所述集群包括引擎服务单元和模型推理服务单元；所述引擎服务单元设置为将发布至目标环境库的大语言模型应用信息按照设定时长加载到所述模型推理服务单元中；所述模型推理服务单元设置为根据所述大语言模型应用信息得到大语言模型的服务信息，并将大语言模型的服务信息进行注册操作，得到大语言模型的注册信息。

其中，设定时长可以是预先设定好的时长。示例性的，设定时长可以是 2 分钟或者 3 分钟，本实施例中可以根据实际需求进行设定。模型推理服务单元中可以包括大语言模型A、大语言模型B以及大语言模型C等多个大语言模型。

本实施例中的各个集群中可以包括引擎服务单元，引擎服务单元设置为将通过配置模块的云端页面发布到目标环境库的大语言模型应用信息按照设定时长加载到模型推理服务单元中，以使得模型推理服务单元将根据大语言模型应用信息获取大语言模型的服务信息，并将大语言模型的服务信息注册到 Consul 注册中心中，从而得到大语言模型的注册信息。

本实施例中通过这样的设置，可以将大语言模型的服务信息进行注册，以便于实现大语言模型的应用推理服务。

在本实施例中，可选的，所述集群还包括应用调度服务单元；所述引擎服务单元设置为将大语言模型应用信息按照设定时长加载到所述应用调度服务单元中；所述应用调度服务单元解析所述大语言模型应用信息，以根据应用信息和注册信息调用模型推理服务单元。

本实施例中的引擎服务单元设置为将通过配置模块的云端页面发布到集群的模型信息得到的目标环境库的大语言模型应用信息按照设定时长加载到应用调度服务中，本实施例中的应用调度服务单元可以读取大语言模型的应用信息的各项参数并进行解析验证，从而确定确保大语言模型信息的准确性，根据应用信息和注册信息调用满足应用需求的模型推理服务单元，并对外提供大语言模型的推理服务。

示例性的，本实施例中应用调度服务单元在注册信息中检索与解析出与应用信息相匹配的条目；其中，注册信息通常存储了可用模型的详细信息，包括模型名称、唯一标识符、部署位置、所需资源等。然后将根据配置信息中的模型名称或标识符，在注册信息中搜索对应的模型条目，若能找到对应的模型条目，则验证该模型具有可用性，还可检查模型是否处于在线状态、是否有足够的资源来支持推理服务等。

本实施例中通过这样的设置，可以通过集群内引擎服务单元、应用调度服务单元和模型推理服务单元对大语言模型应用信息和注册信息进行解析，并调用符合使用需求的模型推理服务单元，提高了大语言模型的应用推理功能的适配性。

在本实施例中，可选的，所述模型推理服务单元还设置为提供大语言模型应用推理服务的接口，以使得用户通过接口调用所述大语言模型进行使用。

本实施例中大语言模型应用推理服务的接口可以是一个接口，还可以是可调用的地址。本实施例中的模型推理服务单元还对外部提供大语言模型应用推理服务的接口，以便于用户基于接口调用大语言模型进行使用推理功能。示例性的，当用户在使用大语言模型时，通过大语言模型应用推理服务的接口发出对应的问题请求，经过大语言模型的应用推理服务进行处理后返回对应的处理结果。本实施例中通过这样的设置，在部署得到满足应用需求的大语言模型之后，可以通过设置对应的外部接口，以便于满足用户的使用需求。

在本实施例中，可选的，所述集群还包括日志推送单元；所述日志推送单元设置为将所述模型推理服务单元和应用调度服务单元的执行记录推送至消息队列中。

其中，执行记录可以认为是模型推理服务单元和应用调度服务单元的执行使用记录情况。在本实施例中，每个集群中都可以包括对应的日志推送功能，在调用大语言模型对应的模型推理服务单元和应用服务单元时，就会对应的执行使用记录信息，并将该执行使用记录信息推送至日志的消息队列中。

本实施例中通过这样的设置，可以通过日志推送功能自动将服务的调用记录推送至日志解析单元中进行解析，提高了集群中日志记录的可靠性。

在本实施例中，可选的，所述集群还包括日志解析单元；所述日志解析单元设置为对所述消息队列中的执行记录进行解析并保存至所述集群的日志库中，以供所述日志查询模块查看。

在本实施例中，每个集群中都可以包括对应的日志解析单元功能，本实施例中日志解析单元可以将消息队列 RabbitMQ 中的日志消息解析保存到该集群日志库中，以便于供日志查询模块进行查看。

本实施例中可以通过日志解析单元对各个集群中的日志信息进行解析并存储至对应的日志存储库中，以便于响应日志查询请求进行查询对应的服务调用记录。

本申请实施例的技术方案，通过配置模块设置为响应于用户触发操作，基于所述用户触发操作得到配置信息，并将所述配置信息发送到至少一个集群中；针对每个集群，所述集群设置为接收所述配置信息，并对所述配置信息进行配置处理后得到大语言模型；其中，所述配置信息包括大语言模型的部署信息以及模型信息；所述日志查询模块设置为接收查询请求，以基于所述查询请求查询每个集群得到所述大语言模型的调用记录信息。本系统通过对多集群的管理操作，实现大语言模型的多场景推理服务应用的灵活性和适配性。

图 2 是根据本申请实施例提供的一种大语言模型的管理系统的结构示例图。如图 2 所示，本申请实施例中的系统整体设计架构主要分为两部分：第一部分包括云端配置、日志查询模块和集群 A 和集群 B，第二部分包括各集群内所包

含的引擎服务单元。如图 2 所示，集群 A 中包括测试环境库和生产环境库；集群 B 中包括生产环境库。

示例性的，本实施例中云端配置可以包括：模型仓库单元、应用管理单元以及调用明细单元。模型仓库单元设置为从基于用户在云端页面上的触发操作将大语言模型上传保存至云端；应用管理单元设置为基于用户在云端页面上的触发操作选择对应的集群和环境将大语言模型应用部署到指定环境和集群；调用明细单元设置为将不同集群内的调用日志展示在云端页面。集群内引擎服务单元包括：引擎服务部分、日志推送及存储部分。引擎服务部分将云端页面发布到指定环境库的大语言模型应用信息定时加载到应用调度服务单元和模型推理服务单元中，模型推理服务单元根据大语言模型的应用信息得到服务信息，将服务信息注册到 Consul 注册中心中，应用调度服务单元设置为解析大语言模型应用信息，根据应用信息和注册信息调用模型推理服务单元，并对外提供大语言模型应用推理服务。日志推送及存储部分，将模型推理服务单元和应用调度服务单元的执行记录推送到 RabbitMQ 中，日志解析部分将 RabbitMQ 中的日志消息解析保存到集群日志库中，供日志查询模块查看。其中，模型推理服务单元中可以包括大语言模型 A、大语言模型 B 以及大语言模型 C。

本实施例中模型的配置信息和大语言模型的模型文件都可以通过云端页面进行配置；其中，模型仓库单元会管理整个页面的整个模型的关系，可以包括模型的名称、类型，所放位置等信息。通过将大语言模型的配置信息部署到指定环境和集群；每个集群内引擎服务部分加载大语言模型的配置信息，大语言模型应用在对应的集群和环境内生效，可以对外提供大语言模型应用服务；外部调用大语言模型应用的记录，通过日志推送及存储部分将模型推理服务单元和应用调度服务单元的的执行记录推送至对应的集群内，在云端调用明细单元中查看各集群日志记录详情。

本申请实施例在大语言模型部署时，先选择环境，再选择部署集群的方式，确定大语言模型最终部署的集群和环境，通过数据操作服务解决部署时的页面配置服务和各集群网络隔离的问题，通过各集群日志分别采集的方式解决各隔离集群调用日志反显的问题。本实施例通过云端页面配置的大语言模型应用先将金融风控场景的大语言模型发布到测试库，由大语言模型推理引擎服务进行

加载，测试验证通过后再发布生产库将大语言模型服务应用于生产，保证了推理服务的可靠性。本实施例中通过云端页面服务与集群间通过数据操作服务进行关联，将云端页面大语言模型应用相关配置同步到发布库，解决了云端配置页面与推理引擎集群间网络隔问题。本实施例中将环境库与集群混合部署，集群中存在环境库，但是又不需要具备完整的环境库，将多集群的隔离与多环境部署策略相结合，解决环境间资源冲突问题，实现了统一的多集群多环境的管理方式。

应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本申请的技术方案所期望的结果，本文在此不进行限制。

权 利 要 求 书

1、一种大语言模型的管理系统，包括配置模块、至少一个集群以及日志查询模块；所述配置模块与至少一个集群通过数据操作服务进行通信连接，所述至少一个集群与所述日志查询模块通过网络通信连接；

所述配置模块设置为响应于用户触发操作，基于所述用户触发操作得到配置信息，并将所述配置信息发送到至少一个集群中；

针对每个集群，所述集群设置为接收所述配置信息，并对所述配置信息进行配置处理后得到与所述配置信息对应的大语言模型；其中，所述配置信息包括大语言模型的部署信息以及模型信息；

所述日志查询模块设置为接收查询请求，基于所述查询请求查询所述集群得到所述大语言模型的调用记录信息。

2、根据权利要求 1 所述的系统，其中，所述配置模块包括模型仓库单元和应用管理单元；

所述模型仓库单元设置为响应所述用户触发操作提供多种功能类型的大语言模型；

所述应用管理单元设置为响应所述用户触发操作选择对应的部署信息。

3、根据权利要求 1 所述的系统，其中，所述集群包括目标环境库；

所述配置模块设置为将所述大语言模型部署至所述大语言模型对应的集群以及目标环境库中。

4、根据权利要求 3 所述的系统，其中，所述目标环境库包括如下至少之一：生产环境库或测试环境库。

5、根据权利要求 4 所述的系统，其中，在所述目标环境库包括所述生产环境库和所述测试环境库的情况下，所述配置模块设置为将大语言模型发布至所述测试环境库进行测试验证，并将测试验证通过的大语言模型部署至所述生产环境库进行生产使用。

6、根据权利要求 1 所述的系统，其中，所述集群包括引擎服务单元和模型推理服务单元；

所述引擎服务单元设置为将发布至目标环境库的大语言模型应用信息按照设定时长加载到所述模型推理服务单元中；

所述模型推理服务单元设置为根据所述大语言模型应用信息得到大语言模

型的服务信息，并将大语言模型的服务信息进行注册操作，得到大语言模型的注册信息。

7、根据权利要求 6 所述的系统，其中，所述集群还包括应用调度服务单元；

所述引擎服务单元还设置为将所述大语言模型应用信息按照设定时长加载到所述应用调度服务单元中；

所述应用调度服务单元设置为解析所述大语言模型应用信息，以根据所述应用信息和所述注册信息调用所述模型推理服务单元。

8、根据权利要求 6 所述的系统，其中，所述模型推理服务单元还设置为提供大语言模型应用推理服务的接口，以使得用户通过接口调用所述大语言模型进行使用。

9、根据权利要求 7 所述的系统，其中，所述集群还包括日志推送单元；

所述日志推送单元设置为将所述模型推理服务单元和所述应用调度服务单元的执行记录推送至消息队列中。

10、根据权利要求 9 所述的系统，其中，所述集群还包括日志解析单元；

所述日志解析单元设置为对所述消息队列中的执行记录进行解析并保存至所述集群的日志库中，以供所述日志查询模块查看。