MANAGEMENT SYSTEM FOR LARGE LANGUAGE MODELS

TECHNICAL FIELD

[0001] The present application relates to the field of artificial intelligence technology, for example, to a management system for large language models.

BACKGROUND

[0002] Since the automotive finance business involves multiple scenarios and data sources, to ensure the security of financial data, inference services often need to be collaboratively managed across at least one cluster and environment.

[0003] However, the current collaborative management technology has great complexity in communication and resource scheduling between multiple clusters, making it difficult to achieve efficient resource utilization and inference performance assurance. Moreover, the isolation and deployment strategies of multiple environments lack a unified management solution, which likely leads to resource conflicts between environments and unreliable inference results.

SUMMARY

[0004] A management system for large language models is provided according to the present application, the management system for large language models includes a configuration module, at least one cluster and a log query module. The configuration module establishes a communication connection with at least one cluster through a data operation service, and the at least one cluster establishes a communication connection with the log query module through a network.

[0005] The configuration module is configured to, in response to a user trigger operation, obtain configuration information based on the user trigger operation, and send the configuration

information to at least one cluster.

[0006] For each cluster of the at least one cluster, the cluster is configured to receive the configuration information, and obtain a large language model corresponding to the configuration information after configuration processing is performed on the configuration information. The configuration information includes deployment information and model information of a large language model.

[0007] The log query module is configured to receive a query request and query the clusters based on the query request to obtain call record information of the large language model.

[0008] Optionally, the configuration module includes a model warehouse unit and an application management unit.

[0009] The model warehouse unit is configured to provide large language models of multiple functional types in response to the user trigger operation.

[0010] The application management unit is configured to select corresponding deployment information in response to the user trigger operation.

[0011] Optionally, the cluster includes a target environment library.

[0012] The configuration module is configured to deploy the large language model to a cluster and a target environment library corresponding to the large language model.

[0013] Optionally, the target environment library includes at least one of: a production environment library or a test environment library.

[0014] Optionally, the configuration module is configured to release a large language model to the test environment library for test verification, and deploy the large language model after passed the test verification to the production environment library for production use.

[0015] Optionally, the cluster includes an engine service unit and a model inference service unit.

[0016] The engine service unit is configured to load application information of a large language model released to a target environment library into the model inference service unit according to

a set duration.

[0017] The model inference service unit is configured to obtain service information of the large language model according to the application information of the large language model, and register the service information of the large language model to obtain registration information of the large language model.

[0018] Optionally, the cluster includes an application scheduling service unit.

[0019] The engine service unit is configured to load the application information of the large language model into the application scheduling service unit according to a set duration.

[0020] The application scheduling service unit is configured to parse the application information of the large language model to call the model inference service unit according to the application information and the registration information.

[0021] Optionally, the model inference service unit is further configured to provide an interface for an application inference service of a large language model, to allow a user to call the large language model through the interface for use.

[0022] Optionally, the cluster further includes a log push unit.

[0023] The log push unit is configured to push execution records of the model inference service unit and the application scheduling service unit to a message queue.

[0024] Optionally, the cluster further includes a log parsing unit.

[0025] The log parsing unit is configured to parse execution records in the message queue and save parsed results to a log library of the cluster for viewing by the log query module.

BRIEF DESCRIPTION OF DRAWINGS

[0026] FIG. 1 is a schematic structural diagram of a management system for large language models provided according to an embodiment of the present application; and

[0027] FIG. 2 is a structural diagram of a management system for large language models

provided according to an embodiment of the present application.

DETAILED DESCRIPTION

[0028] The terms "target", "original", etc., in the specification and claims of the present application and the above-mentioned drawings are used to distinguish similar objects and are not necessarily used to describe a specific order or sequence. It should be understood that the data used in this way can be interchanged where appropriate, so that the embodiments of the present application described herein can be implemented in an order other than those illustrated or described herein. In addition, the terms "including" and "having" and any variations thereof are intended to cover non-exclusive inclusions, for example, the processes, methods, systems, products or devices, in addition to including a series of steps or units shown in the embodiments of the present application, may further include other steps or units that are not listed, or include other steps or units inherent to these processes, methods, systems, products or devices.

[0029] FIG. 1 is a structural diagram of a management system for large language models provided according to an embodiment of the present application. As shown in FIG. 1, the management system for large language models includes a configuration module 110, at least one cluster 120, and a log query module 130. The configuration module establishes a communication connection with at least one cluster through a data operation service, and the at least one cluster establishes a communication connection with the log query module through a network. The data operation service may refer to a service being responsible for processing data transmission and interaction between large language models and clusters based on a set communication mechanism or service. In this embodiment, the data operation service may refer to realizing the communication between the configuration module and at least one cluster through a set communication interface. Exemplarily, the data operation service may be an operation (OP) service. The configuration module in this embodiment may be presented to the user by

configuring through a cloud page, so that the user selects the configuration content through a trigger operation to obtain the corresponding configuration information, such as the deployment location for configuring the large language model, the specific objects for the deployment, and the storage location. The large language model in this embodiment can be directed to a large language model for the automotive financial risk control scenario.

[0030] The configuration module 110 is configured to, in response to a user trigger operation, obtain configuration information based on the user trigger operation, and send the configuration information to at least one cluster.

[0031] The configuration information includes the deployment information and model information of the large language model. The deployment information may include information such as the deployment objects and the specific deployment location. The deployment objects may be objects such as a deployment specified environment and cluster. The deployment location may be the location where the large language model needs to be deployed in the end. The model information may refer to the storage location information of the large language model. In this embodiment, the at least one cluster may include a cluster A, a cluster B, a cluster C... a cluster N, etc., and the specific number is not limited.

[0032] In this embodiment, the user selects contents such as the deployment information and model information of the large language model through a trigger operation, and in response to the trigger operation and according to the trigger operation, the configuration module obtains the configuration information selected by the user, and sends the configuration information to at least one cluster. In this embodiment, the management system for large language models adopts a unified multi-cluster and multi-environment management method, specifically deploying the environment libraries and the clusters in a mixed manner, so that there is an environment library in a cluster, but it is not necessary to have a complete environment library in the cluster, thereby combining the isolation of multiple clusters with the multi-environment deployment strategy to

solve the problem of resource conflicts between environments. It can be understood that at least one cluster in this embodiment may include one or more specified environment libraries, so that when the large language model is deployed, the corresponding cluster and the specified environment library contained therein can be selected. The environment libraries may include, but are not limited to, a test environment library and a production environment library.

[0033] In this embodiment, when the large language model is deployed based on the configuration module, the environment libraries can be selected first, and then the clusters into which the large language model is to be deployed can be selected to determine the cluster objects and environment objects to which the large language model is finally deployed. In this embodiment, the page configuration service and the isolation of clusters through a network during deployment can be solved by the data operation service, and the problem of data binding in calling logs of isolated clusters can be solved by collecting the logs of the clusters separately. In order to ensure the reliability of the inference results as much as possible, the large language model in the financial risk control scenario needs to be released and deployed into the test environment for verification first, and then the same large language model service application is deployed into the production environment for use.

[0034] For each cluster 120, the cluster 120 is configured to receive the configuration information, and obtain a large language model corresponding to the configuration information after configuration processing is performed on the configuration information.

[0035] After configuration processing is performed on the configuration information based on the included engine service, the cluster can obtain the corresponding large language model. The engine service can be a software service for processing tasks related to the large language model. The engine service included in the cluster in this embodiment can generally include multiple functional services for processing different aspects of the large language model. Exemplarily, the engine service may include training service, inference service, optimization service, etc., and each

service is responsible for a specific task.

[0036] In this embodiment, after receiving the configuration information corresponding to the large language model sent by the configuration module, the clusters can first process the relevant application information of the large language model through the environment libraries according to the different environment libraries included in the clusters, and then process other services such as the engine service, so as to obtain the configured large language model.

[0037] The log query module 130 is configured to receive a query request and query the clusters based on the query request to obtain call record information of the large language model.

[0038] The query request can be a request issued by the user to query the usage of a large language model in clusters after each cluster obtains the corresponding large language model. In this embodiment, each cluster has its own log push function. Each cluster generates a call record each time a different service in the cluster is called, and then the call record is pushed to a call detail unit through the log push function, so that the log query module can query it. The configuration module includes a call detail unit for presenting the call records generated by different clusters on the cloud page. In this embodiment, in response to receiving a log query request issued by a user, the corresponding call record information and other usage information in at least one corresponding cluster is queried through the query request.

[0039] In this embodiment, optionally, the configuration module includes a model warehouse unit and an application management unit. The model warehouse unit is configured to provide large language models of multiple functional types in response to user trigger operations. The application management unit is configured to select corresponding deployment information in response to the user trigger operation.

[0040] The configuration module of the management system for large language models in this embodiment includes a model warehouse unit and an application management unit. The model warehouse unit may include large language models of various functional types, and is configured

to upload and save a large language model to the cloud in response to a user trigger operation. The application management unit is configured to select a corresponding cluster and environment in response to a user trigger operation, so as to deploy the large language model application to the corresponding cluster and environment.

[0041] In this embodiment, information such as the model information and a model file of the large language model can be configured through the cloud page, so that the model configuration that meets the user's needs can be flexibly determined.

[0042] In this embodiment, optionally, the cluster includes a target environment library. The configuration module is configured to deploy the large language model to a cluster and a target environment library corresponding to the large language model.

[0043] The target environment library includes a production environment library and/or a test environment library. The test environment library can be used to perform application test operations on the large language model. The production environment library can be used to produce large language model service applications. It can be understood that different clusters can include different target environment libraries. For example, a cluster A can include a production environment library; a cluster B can include a test environment library; and a cluster C can include a production environment library and a test environment library. In this embodiment, the target environment library contained in a cluster can be determined according to the specific orientation settings and other information of the cluster.

[0044] In this embodiment, the large language model can be deployed to at least one corresponding cluster and the target environment library corresponding to the cluster through the configuration module. In this embodiment, through such settings, multiple clusters and multiple environments can be managed to provide users with different types of deployment settings to meet user requirements.

[0045] In this embodiment, optionally, the configuration module is configured to release a large

language model to the test environment library for test verification, and deploy the large language model after passed the test verification to the production environment library for production use.

[0046] In this embodiment, the configuration module can send a large language model to the corresponding test environment library to test and verify the application function of the large language model, and release a large language model that passes the test verification to the corresponding production environment library, so as to apply the large language model service to production use through the production environment library.

[0047] In this embodiment, through such a setting, it is necessary to first release and deploy a large language model service application into the test environment for verification, and then deploy the same large language model service application to the production environment for use, thereby ensuring the reliability of inference results of the large language model in the financial risk control scenario.

[0048] In this embodiment, optionally, the cluster includes an engine service unit and a model inference service unit. The engine service unit is configured to load application information of a large language model released to a target environment library into the model inference service unit according to a set duration. The model inference service unit is configured to obtain service information of the large language model according to the application information of the large language model, and register the service information of the large language model to obtain registration information of the large language model.

[0049] The set duration can be a pre-set duration. Exemplarily, the set duration can be 2 minutes or 3 minutes, which can be set according to practical requirements in this embodiment. The model inference service unit may include multiple large language models, such as a large language model A, a large language model B, and a large language model C.

[0050] In this embodiment, each cluster may include an engine service unit, and the engine service unit is configured to load the application information of the large language model, released

to the target environment library through the cloud page of the configuration module, into the model inference service unit according to the set duration, so that the model inference service unit will obtain the service information of the large language model according to the application information of the large language model, and register the service information of the large language model in the Consul registration center, thereby obtaining the registration information of the large language model.

[0051] In this embodiment, through such a setting, the service information of the large language model can be registered to facilitate the implementation of an application inference service of the large language model.

[0052] In this embodiment, optionally, the cluster further includes an application scheduling service unit. The engine service unit is configured to load the application information of the large language model into the application scheduling service unit according to a set duration. The application scheduling service unit is configured to parse the application information of the large language model to call the model inference service unit according to the application information and the registration information.

[0053] The engine service unit in this embodiment is configured to load the application information of the large language model of the target environment library, obtained by the model information released to the cluster through the cloud page of the configuration module, into the application scheduling service unit according to the set duration. The application scheduling service unit in this embodiment can read various parameters of the application information of the large language model to parse and verify, so as to ensure the accuracy of the large language model information, call the model inference service unit that meets the application requirements according to the application information and the registration information, and provide the inference service of the large language model to the external clients.

[0054] Exemplarily, in this embodiment, an entry matching the application information is

retrieved in the registration information and parsed by the application scheduling service unit. The registration information generally stores detailed information of an available model, including a model name, a unique identifier, a deployment location, a required resource, etc. Then, according to the model name or identifier in the configuration information, the corresponding model entry will be searched in the registration information. If the corresponding model entry can be found, it verifies that the model is available, and it may further be checked whether the model is online and whether there are sufficient resources to support the inference service.

[0055] In this embodiment, through such a setting, the application information and registration information of the large language model can be parsed through the engine service unit, application scheduling service unit and model inference service unit in the cluster, and the model inference service unit that meets the use requirements can be called, thereby improving the adaptability of the application inference function of the large language model.

[0056] In this embodiment, optionally, the model inference service unit is further configured to provide an interface for an application inference service of a large language model, to allow a user to call the large language model through the interface for use.

[0057] In this embodiment, the interface of the application inference service of the large language model can be an interface or a callable address. The model inference service unit in this embodiment further provides an interface for the application inference service of the large language model to the outside, so that the user can call the large language model based on the interface to use the inference function. Exemplarily, when the user is using the large language model, the corresponding question request is issued through the interface of the application inference service of the large language model, and the corresponding processing result is returned after the question request is processed by the application inference service of the large language model. In this embodiment, through such a setting, after the large language model that meets the application requirements is deployed, the corresponding external interface can be set to meet the

user's usage requirements.

[0058] In this embodiment, optionally, the cluster further includes a log push unit. The log push unit is configured to push execution records of the model inference service unit and the application scheduling service unit to a message queue.

[0059] The execution records can be considered as the execution and use records of the model inference service unit and the application scheduling service unit. In this embodiment, each cluster can include a corresponding log push function. When the model inference service unit and the application scheduling service unit corresponding to the large language model are called, the corresponding execution and use record information will be generated, and the execution and use record information will be pushed to the message queue of the log.

[0060] In this embodiment, through such a setting, the service call record can be automatically pushed through the log push function to a log parsing unit for parsing, thereby improving the reliability of the log records in the cluster.

[0061] In this embodiment, optionally, the cluster further includes a log parsing unit. The log parsing unit is configured to parse execution records in the message queue and save parsed results to a log library of the cluster for viewing by the log query module.

[0062] In this embodiment, each cluster may include a corresponding log parsing unit function. In this embodiment, the log parsing unit may parse the log message in the message queue RabbitMQ and save parsed results to the log library of the cluster for viewing by the log query module.

[0063] In this embodiment, the log information in each cluster may be parsed by the log parsing unit and then stored in the corresponding log storage library, so as to, in response to the log query request, query the corresponding service call records.

[0064] In the technical solution of the embodiment of the present application, the configuration module is configured to, in response to a user trigger operation, obtain configuration information

based on the user trigger operation, and send the configuration information to at least one cluster; for each cluster, the cluster is configured to receive the configuration information, and obtain a large language model after configuration processing is performed on the configuration information. The configuration information includes deployment information and model information of a large language model; the log query module is configured to receive a query request, so as to query each cluster based on the query request to obtain the call record information of the large language model. Therefore, the system realizes the flexibility and adaptability of the multi-scenario inference service application of the large language model through the management operation for multiple clusters.

[0065] FIG. 2 is a structural example diagram of a management system for large language models provided according to an embodiment of the present application. As shown in FIG. 2, the overall design architecture of the system in the embodiment of the present application is mainly divided into two parts: a first part includes a cloud configuration, log query modules, a cluster A and a cluster B, and a second part includes engine service units contained in the cluster. As shown in FIG. 2, the cluster A includes a test environment library and a production environment library; and the cluster B includes a production environment library.

[0066] Exemplarily, the cloud configuration in this embodiment may include: a model warehouse unit, an application management unit, and a call detail unit. The model warehouse unit is configured to upload and save a large language model to the cloud based on the trigger operation of the user on the cloud page; the application management unit is configured to select a corresponding cluster and environment based on the trigger operation of the user on the cloud page to deploy the large language model application to the specified environment and cluster; and the call detail unit is configured to present the call logs in different clusters on the cloud page. The engine service unit in the cluster includes: an engine service section and a log push and storage section. The engine service section loads application information of the large language

model, released to a specified environment library through the cloud page, into an application scheduling service unit and a model inference service unit according to a set duration. The model inference service unit obtains service information according to the application information of the large language model, and registers the service information in the Consul registration center. The application scheduling service unit is set to parse the application information of the large language model, call the model inference service unit according to the application information and registration information, and provide an application inference service of the large language model to the outside. The log push and storage section pushes execution records of the model inference service unit and the application scheduling service unit to RabbitMQ, and the log parsing section parses the log message in RabbitMQ and saves parsed results into the cluster log library for the log query module to view. The model inference service unit may include a large language model A, a large language model B and a large language model C.

[0067] In this embodiment, configuration information of models and model files of the large language models can all be configured through the cloud page. The model warehouse unit manages the relationships of the entire model of the entire page, which may include the name, type, location and other information of the model. By deploying the configuration information of the large language model to the specified environment and cluster; the engine service section in each cluster loads the configuration information of the large language model, and the large language model application takes effect in the corresponding cluster and environment, and can provide application services of the large language model to the outside. When the records of the large language model applications are called externally, the execution records of the model inference service unit and the application scheduling service unit are pushed by the log push and storage section to the corresponding cluster, and the cluster log record details can be viewed in the cloud call detail unit.

[0068] In the embodiment of the present application, when the large language model is to be

deployed, the environment is selected first, then the way to deploy the cluster is selected, the cluster and environment where the large language model is finally deployed is determined, the problems of page configuration service and isolation of clusters through a network during deployment are solved through data operation services, and the problem of data binding of call logs of isolated clusters is solved by collecting logs of the clusters separately. In this embodiment, the large language model application configured through the cloud page first releases the large language model of the financial risk control scenario to the test library, which is loaded by the inference engine service of the large language model, and then releases the large language model after passed the test verification into the production library to apply the large language model service to production, which ensures the reliability of the inference service. In this embodiment, the cloud page service is associated with the clusters through the data operation service, and the cloud page large language model application-related configuration is synchronized to the release library, which solves the problem of isolation through a network between the cloud configuration page and the inference engine cluster. In this embodiment, the environment library and the cluster are deployed in a mixed manner, the cluster has an environment library, but it does not need to have a complete environment library, so that the isolation of multiple clusters is combined with the multi-environment deployment strategy, thereby, the problem of resource conflict between environments is solved and a unified multi-cluster and multi-environment management method is realized.

[0069] It should be understood that the various forms of processes shown above can be used with steps reordered, added, or deleted. For example, the steps recorded in this application can be executed in parallel, sequentially, or in different orders, as long as the expected results of the technical solution of this application can be achieved, this application does not limit this here.

Claims

1. A management system for large language models, comprising a configuration module, at least one cluster and a log query module;

wherein the configuration module establishes a communication connection with at least one cluster through a data operation service, and the at least one cluster establishes a communication connection with the log query module through a network;

wherein the configuration module is configured to, in response to a user trigger operation, obtain configuration information based on the user trigger operation, and send the configuration information to the at least one cluster;

for each cluster of the at least one cluster, the cluster is configured to receive the configuration information, and obtain a large language model corresponding to the configuration information after configuration processing is performed on the configuration information;

wherein the configuration information comprises deployment information and model information of the large language model; and

wherein the log query module is configured to receive a query request, and query the at least one cluster based on the query request to obtain call record information of the large language model.

2. The management system for large language models according to claim 1, wherein the configuration module comprises a model warehouse unit and an application management unit;

the model warehouse unit is configured to provide large language models of a plurality of functional types in response to the user trigger operation; and

the application management unit is configured to select corresponding deployment information in response to the user trigger operation.

3. The management system for large language models according to claim 1, wherein the cluster comprises a target environment library;

the configuration module is configured to deploy the large language model to a cluster and a target environment library corresponding to the large language model.

4. The management system for large language models according to claim 3, wherein the target environment library comprises at least one of: a production environment library or a test environment library.

5. The management system for large language models according to claim 4, wherein, in a case where the target environment library comprises the production environment library and the test environment library, the configuration module is configured to release a large language model to the test environment library for test verification, and deploy the large language model after passed the test verification to the production environment library for production use.

6. The management system for large language models according to claim 1, wherein the cluster comprises an engine service unit and a model inference service unit;

the engine service unit is configured to load application information of a large language model released to a target environment library into the model inference service unit according to a set duration; and

the model inference service unit is configured to: obtain service information of the large language model according to the application information of the large language model, and register the service information of the large language model to obtain registration information of the large language model.

7. The management system for large language models according to claim 6, wherein the cluster further comprises an application scheduling service unit;

the engine service unit is further configured to load the application information of the large language model into the application scheduling service unit according to the set duration; and

the application scheduling service unit is configured to parse the application information of the large language model to call the model inference service unit according to the application information and the registration information.

8. The management system for large language models according to claim 6, wherein the model inference service unit is further configured to provide an interface for an application inference service of the large language model, to allow a user to call the large language model through the interface for use.

9. The management system for large language models according to claim 7, wherein the cluster further comprises a log push unit; and

the log push unit is configured to push execution records of the model inference service unit and the application scheduling service unit to a message queue.

10. The management system for large language models according to claim 9, wherein the cluster further comprises a log parsing unit; and

the log parsing unit is configured to parse execution records in the message queue and save parsed results to a log library of the cluster for viewing by the log query module.