MANAGEMENT SYSTEM FOR LARGE LANGUAGE MODELS

ABSTRACT

A management system for large language models includes a configuration module, at least one cluster and a log query module. 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 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 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 and query the clusters based on the query request to obtain call record information of the large language model.

Fig. 01