

REQUEST SCHEDULING METHOD AND DEVICE

ABSTRACT

The present invention discloses a request scheduling method and a device. The method includes: A first device receives a first request from a first user, where the first request is a request
5 for a first service, and the first request includes an identifier of the first service; the first device schedules the first request to a second device; the second device sends a second request to the first device, where the second request includes the identifier of the first service and first output information, and the second request is associated with the first request; when the first output information does not meet a specified condition, the first device schedules the second request to a
10 third device, where the third device is different from the second device; the third device sends a third request to the first device, where the third request includes the identifier of the first service and second output information, and the third request is associated with the second request; and when the second output information does not meet the specified condition, the first device schedules the third request to the third device, so that a throughput of a service system in a large-
15 scale language model service deployment can be improved.

FIG. 8