

LARGE LANGUAGE MODEL INFERENCE METHOD AND APPARATUS

ABSTRACT

Embodiments of this application disclose a large language model inference method and apparatus, to reduce bandwidth consumption of large language model inference. The method in
5 embodiments of this application includes: A compute device obtains input data of a large language
model running on a plurality of processors, where the input data includes input text of a user and
reference information generated based on the input text, and the reference information includes
domain-specific knowledge corresponding to the input text and time-sensitive information
corresponding to the input text; performs matching with historical key-value cache data in a shared
10 storage pool based on the input data, to determine key-value cache data corresponding to the input
data, where the historical key-value cache data indicates intermediate data generated by the large
language model when processing different historical input data, the shared storage pool is used to
store the historical key-value cache data generated by the plurality of processors, and the different
historical input data includes a historical input text and reference information corresponding to the
15 historical input text; and generates, based on the key-value cache data, an output result
corresponding to the input data.

(FIG.2)