MODEL INFERENCE METHOD AND SYSTEM, AND COMPUTE DEVICE

ABSTRACT

A model inference method and system are provided. The method includes: receiving text prompt information, where the text prompt information is used to generate text information; determining at least one draft verification model from a draft verification model pool; determining at least one corresponding draft generation model from a draft generation model pool; obtaining, from a draft pool, a target draft related to the text prompt information, where the draft pool stores a plurality of drafts, and the plurality of drafts include a draft generated by the at least one draft generation model; verifying the target draft by using the at least one draft verification model, to obtain a verification result; and generating the text information based on the verification result, and feeding back the text information. The method can improve model inference efficiency.

Fig. 4