

TEXT GENERATION METHODS AND APPARATUSES, STORAGE MEDIUM DEVICES, AND PROGRAM PRODUCTS

ABSTRACT

This specification provides a text generation method and apparatus, a storage medium device,
5 and a program product, which are applied to an LLM. In each round of iteration of the LLM, the
following operations can be performed: A first text sequence following a current text sequence is
obtained by using a speculative decoding method, and a plurality of candidate sequences are
formed. Logical blocks are allocated to text units in the plurality of candidate sequences in a key-
value cache, and the logical blocks are mapped to physical blocks. In a process of mapping to a
10 physical block, a plurality of logical blocks allocated to the same text unit in the plurality of
candidate sequences obtained through speculative decoding are mapped to the same physical block.
Then, a newly generated text unit in the current round of iteration is determined by the LLM by
using attention information stored in the key-value cache. When speculative decoding is combined
with paging management, in a special scenario in which there is the same text unit in a plurality of
15 candidate sequences, the first criterion for mapping a logical block to a physical block is provided,
so that the plurality of candidate sequences can occupy as little graphics memory space as possible,
to improve utilization of the graphics memory space.

FIG. 2

20