

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2025/152397 A1

(43) 国际公布日
2025 年 7 月 24 日 (24.07.2025)

(51) 国际专利分类号:
G06N 5/04 (2023.01)

(21) 国际申请号: PCT/CN2024/109727

(22) 国际申请日: 2024 年 8 月 5 日 (05.08.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410077095.0 2024 年 1 月 18 日 (18.01.2024) CN
202410223760.2 2024 年 2 月 28 日 (28.02.2024) CN

(71) 申请人: 华为云计算技术有限公司 (HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.)
[CN/CN]; 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。

(72) 发明人: 陈旭升 (CHEN, Xusheng); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心

550025 (CN)。单一舟 (SHAN, Yizhou); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。冯浩 (FENG, Hao); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。张秦 (ZHANG, Qin); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。邓晓阳 (DENG, Xiaoyang); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。但庚元 (DAN, Gengyuan); 中国贵州省贵阳市贵安新区黔中大道交兴功路华为云数据中心, 550025 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦 B 座 18C2、18D、18E、18E2 518001 (CN)。

(54) Title: METHOD AND APPARATUS FOR INFERENCE IN LARGE LANGUAGE MODEL

(54) 发明名称: 一种大语言模型的推理方法以及装置

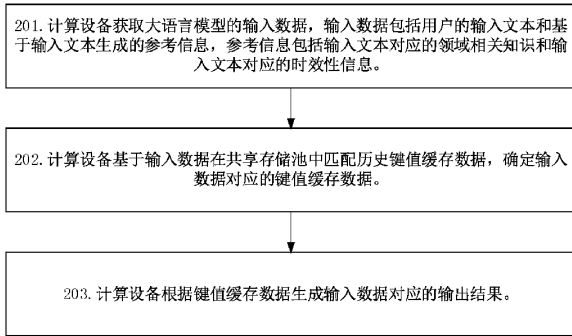


图 2

201 A computing device acquires input data of a large language model, wherein the input data comprises input text of a user and reference information generated on the basis of the input text, and the reference information comprises domain-related knowledge corresponding to the input text and timeliness information corresponding to the input text
202 On the basis of the input data, the computing device matches historical key-value cache data in a shared storage pool, so as to determine key-value cache data corresponding to the input data
203 On the basis of the key-value cache data, the computing device generates an output result corresponding to the input data

(57) Abstract: Disclosed in the embodiments of the present application are a method and apparatus for inference in a large language model, which are used for reducing the bandwidth consumption of inference in a large language model. The method in the embodiments of the present application comprises: a computing device acquiring input data of a large language model running on a plurality of processors, wherein the input data comprises input text of a user and reference information generated on the basis of the input text, and the reference information comprises domain-related knowledge corresponding to the input text and timeliness information corresponding to the input text; on the basis of the input data, matching historical key-value cache data in a shared storage pool, so as to determine key-value cache data corresponding to the input data, wherein the historical key-value cache data is used for indicating intermediate data generated when the large language model processes different historical input data, the shared storage pool is used for storing historical key-value cache data generated by the plurality of processors, and the different historical input data comprises historical input text and reference information corresponding to the historical input text; and on the basis of the key-value cache data, generating an output result corresponding to the input data.

[见续页]

WO 2025/152397 A1

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本申请实施例公开了一种大语言模型的推理方法以及装置, 用于降低大语言模型的推理的带宽消耗。本申请实施例方法包括: 计算设备获取运行在多个处理器上的大语言模型的输入数据, 输入数据包括用户的输入文本和基于输入文本生成的参考信息, 参考信息包括输入文本对应的领域相关知识和输入文本对应的时效性信息。基于输入数据在共享存储池中匹配历史键值缓存数据, 确定输入数据对应的键值缓存数据, 历史键值缓存数据用于指示大语言模型在处理不同历史输入数据时生成的中间数据, 共享存储池用于存储多个处理器生成的历史键值缓存数据, 不同历史输入数据包括历史输入文本和历史输入文本对应的参考信息。根据键值缓存数据生成输入数据对应的输出结果。