

(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 2025/152398 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) **G06F 9/50** (2006.01)
- (21) 国际申请号: PCT/CN2024/109747
- (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:
202410077193.4 2024 年 1 月 18 日 (18.01.2024) CN
202410223757.0 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: REASONING METHOD AND DEVICE FOR LARGE LANGUAGE MODEL

(54) 发明名称: 一种大语言模型的推理方法以及装置

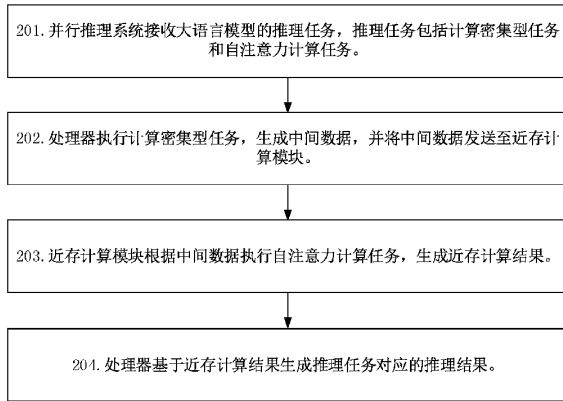


图 2

- 201 A parallel reasoning system receives a reasoning task of a large language model, wherein the reasoning task comprises a compute-intensive task and a self-attention computing task
- 202 A processor executes the compute-intensive task, generates intermediate data, and sends the intermediate data to a near-memory computing module
- 203 The near-memory computing module executes the self-attention computing task on the basis of the intermediate data to generate a near-memory computing result
- 204 On the basis of the near-memory computing result, the processor generates a reasoning result corresponding to the reasoning task

(57) Abstract: Disclosed in embodiments of the present application are a reasoning method and device for a large language model, which are used for reducing the bandwidth consumption of large language model reasoning. The method of the embodiments of the present application comprises: a parallel reasoning system receives a reasoning task of a large language model, wherein the reasoning task comprises a compute-intensive task and a self-attention computing task; a processor executes the compute-intensive task, generates intermediate data, and sends the intermediate data to a near-memory computing module, wherein the compute-intensive task comprises one or more tasks of the large language model: a feedforward neural network computing task, a projection task, and a layer normalization task; the near-memory computing module executes the self-attention computing task on the basis of the intermediate data to generate a near-memory computing result; and on the basis of the near-memory computing result, the processor generates a reasoning result corresponding to the reasoning task.

[见续页]

(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: 并行推理系统接收大语言模型的推理任务, 推理任务包括计算密集型任务和自注意力计算任务。处理器执行计算密集型任务, 生成中间数据, 并将中间数据发送至近存计算模块, 计算密集型任务包括大语言模型的以下一项或多项任务: 前馈神经网络计算任务、投影任务和层归一化任务。近存计算模块根据中间数据执行自注意力计算任务, 生成近存计算结果。处理器基于近存计算结果生成推理任务对应的推理结果。