(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

(43) 国际公布日
2025 年 9 月 12 日 (12.09.2025)

WIPO | PCT

(10) 国际公布号
WO 2025/185312 A1

(51) 国际专利分类号:
G06N 5/04 (2023.01)

(21) 国际申请号: PCT/CN2024/142366

(22) 国际申请日: 2024 年 12 月 25 日 (25.12.2024)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202410263183.X 2024年3月7日 (07.03.2024) CN
202410381622.7 2024年3月29日 (29.03.2024) 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: 段新宇 (**DUAN, Xinyu**); 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)。 夏庆荣 (**XIA, Qingrong**); 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)。 叶子 (**YE, Zi**); 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)。 蒋炜 (**JIANG, Wei**); 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)。 王喆锋 (**WANG, Zhefeng**); 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: 北京龙双利达知识产权代理有限公司 (**LONGSUN LEAD IP LTD.**); 中国北京市海淀区北清路81号院二区3号楼8层801-1室 100094 (CN)。

(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,

(54) **Title:** MODEL INFERENCE METHOD AND SYSTEM, AND COMPUTING DEVICE

(54) 发明名称: 模型推理的方法、系统以及计算设备

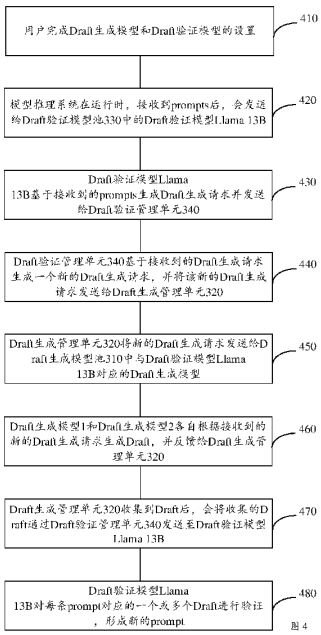


图 4

410 A user completes draft generation model and draft verification model configuration
420 When a model inference system is running, after receiving prompts, the model inference system sends the prompts to a draft verification model Llama 13B in a draft verification model pool 330
430 On the basis of the received prompts, the draft verification model Llama 13B generates a draft generation request and sends same to a draft verification management unit 340
440 On the basis of the received draft generation request, the draft verification management unit 340 generates a new draft generation request, and sends the new draft generation request to a draft generation management unit 320
450 The draft generation management unit 320 sends the new draft generation request to draft generation models in a draft generation model pool 310, the draft generation models corresponding to the draft verification model Llama 13B
460 On the basis of the received new draft generation request, draft generation model 1 and draft generation model 2 each generate a draft and feed back the draft to the draft generation management unit 320
470 Once the draft generation management unit 320 has collected the drafts, the draft generation management unit 320 sends the collected drafts to the draft verification model Llama 13B via the draft verification management unit 340
480 The draft verification model Llama 13B performs verification on one or more drafts corresponding to each prompt, so as to form a new prompt

(57) **Abstract:** A model inference method and system. The method comprises: receiving text prompt information, the text prompt information being used for generating 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; acquiring from a draft pool a target draft related to the text prompt information, a plurality of drafts being stored in the draft pool and comprising a draft generated by the at least one draft generation model; verifying the target draft by means of the at least one draft verification model to obtain

WO 2025/185312 A1

[见续页]

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))。

a verification result; and, on the basis of the verification result, generating text information and feeding back same. The method can improve model inference efficiency.

(57) 摘要: 一种模型推理的方法和系统，方法包括：接收文本提示信息，文本提示信息用于生成文本信息；从Draft验证模型池中确定至少一个Draft验证模型；从Draft生成模型池中确定对应的至少一个Draft生成模型；从Draft池中获取与文本提示信息相关的目标Draft，Draft池中存储有多个Draft，多个Draft包括至少一个Draft生成模型生成的Draft；通过至少一个Draft验证模型对目标Draft进行验证得到验证结果；根据验证结果生成文本信息并反馈。方法能够提高模型推理的效率。