

(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 2025/129623 A1

(43) 国际公布日
2025 年 6 月 26 日 (26.06.2025)

(51) 国际专利分类号:
G06F 16/332 (2019.01)

(21) 国际申请号: PCT/CN2023/140974

(22) 国际申请日: 2023 年 12 月 22 日 (22.12.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 新华三技术有限公司 (NEW H3C TECHNOLOGIES CO., LTD.) [CN/CN]; 中国浙江省杭州市滨江区长河路466号 310052 (CN)。

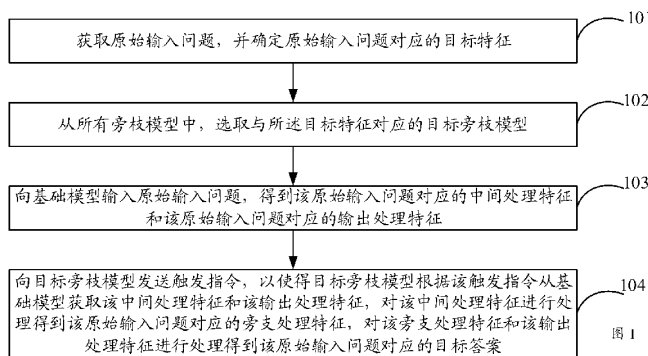
(72) 发明人: 王明辉 (WANG, Minghui); 中国北京市朝阳区广顺南大街8号院1号楼利星行中心A座640室 100102 (CN)。

(74) 代理人: 北京博思佳知识产权代理有限公司 (BEIJING BESTIPR INTELLECTUAL PROPERTY LAW CORPORATION); 中国北京市海淀区上地三街9号嘉华大厦B座409 100085 (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, 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。

(54) Title: DATA PROCESSING METHOD AND APPARATUS BASED ON LARGE LANGUAGE MODEL, AND DEVICE

(54) 发明名称: 一种基于大语言模型的数据处理方法、装置及设备



- 101 Obtain an original input question, and determine a target feature corresponding to the original input question
- 102 Select a target branch model corresponding to the target feature from among all branch models
- 103 Input the original input question to a basic model, so as to obtain an intermediate processing feature corresponding to the original input question and an output processing feature corresponding to the original input question
- 104 Send a trigger instruction to the target branch model, so that the target branch model obtains the intermediate processing feature and the output processing feature from the basic model on the basis of the trigger instruction, processes the intermediate processing feature so as to obtain a branch processing feature corresponding to the original input question, and processes the branch processing feature and the output processing feature so as to obtain a target answer corresponding to the original input question

(57) Abstract: The present application provides a data processing method and apparatus based on a large language model, and a device. The method comprises: obtaining an original input question, and determining a target feature corresponding to the original input question; selecting a target branch model corresponding to the target feature from among all branch models; inputting the original input question to a basic model, so as to obtain an intermediate processing feature corresponding to the original input question and an output processing feature corresponding to the original input question; and sending a trigger instruction to the target branch model, so that the

[见续页]

(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))。

target branch model obtains the intermediate processing feature and the output processing feature from the basic model on the basis of the trigger instruction, processes the intermediate processing feature so as to obtain a branch processing feature corresponding to the original input question, and processes the branch processing feature and the output processing feature so as to obtain a target answer corresponding to the original input question. The technical solution of the present application saves computing resources of a server and time overhead, and thus a target answer can be quickly determined.

(57) 摘要: 本申请提供一种基于大语言模型的数据处理方法、装置及设备, 该方法包括: 获取原始输入问题, 并确定原始输入问题对应的目标特征; 从所有旁枝模型中, 选取与目标特征对应的目标旁枝模型; 向基础模型输入原始输入问题, 得到原始输入问题对应的中间处理特征和原始输入问题对应的输出处理特征; 向目标旁枝模型发送触发指令, 以使得目标旁枝模型根据触发指令从基础模型获取中间处理特征和输出处理特征, 对中间处理特征进行处理得到原始输入问题对应的旁支处理特征, 对旁支处理特征和输出处理特征进行处理得到原始输入问题对应的目标答案。通过本申请的技术方案, 能够节省服务器的计算资源, 并节省时间开销, 能够快速确定出目标答案。