(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
2026 年 9 月 17 日 (17.09.2026)

WIPO | PCT

(10) 国际公布号
WO 2026/188656 A1

(51) 国际专利分类号:
G06F 11/30 (2006.01) *G06N 5/04* (2023.01)
G06N 3/0455 (2023.01) *G06F 16/334* (2025.01)
G06N 3/08 (2023.01)

(21) 国际申请号: PCT/CN2025/096494

(22) 国际申请日: 2025 年 5 月 22 日 (22.05.2025)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202510295398.4 2025年3月13日 (13.03.2025) CN

(71) 申请人: 无限星途科技私人有限公司 (X STAR TECHNOLOGY PTE. LTD.) [SG/SG]; 新加坡, 新达城1号楼, 37-03, 淡马锡林荫道7号 038987 (SG)。

(72) 发明人: 杨青相 (YANG, Qingxiang); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 唐守旭 (TANG, Shouxu); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 张磊 (ZHANG, Lei); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 田立民 (TIAN, Limin); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 刘树祥 (LIU, Shuxiang); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。 刘兵 (LIU, Bing); 中国上海市浦东新区中国(上海)自由贸易试验区浦东大道2123号三层 200135 (CN)。

(74) 代理人: 北京品源专利代理有限公司 (BEYOND ATTORNEYS AT LAW); 中国北京市海淀区莲花池东路39号西金大厦6层 100036 (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,

(54) **Title:** MANAGEMENT SYSTEM FOR LARGE LANGUAGE MODEL

(54) **发明名称:** 大语言模型的管理系统

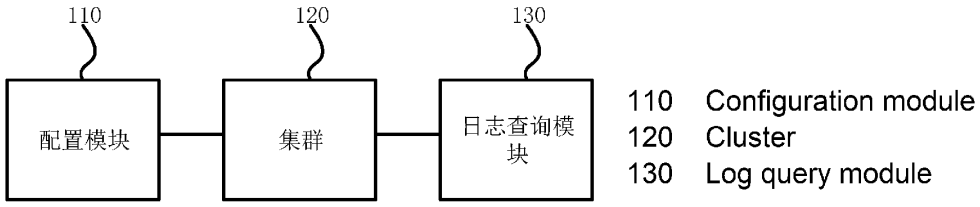


图 1

(57) **Abstract:** A management system for a large language model, comprising a configuration module, at least one cluster, and a log query module. The configuration module is configured to, in response to a user trigger operation, obtain configuration information on the basis of the user trigger operation, and send the configuration information to the at least one cluster; for each cluster, the cluster is configured to receive the configuration information, and perform configuration processing on the configuration information to obtain a large language model corresponding to the configuration information, wherein the configuration information comprises deployment information and model information of the large language model; and the log query module is configured to receive a query request, so as to query the cluster on the basis of the query request to obtain call record information of the large language model.

(57) **摘要:** 一种大语言模型的管理系统，包括配置模块、至少一个集群以及日志查询模块；所述配置模块设置为响应于用户触发操作，基于所述用户触发操作得到配置信息，并将所述配置信息发送到至少一个集群中；针对每个集群，所述集群设置为接收所述配置信息，并对所述配置信息进行配置处理后得到与所述配置信息对应的大语言模型；其中，所述配置信息包括大语言模型的部署信息以及模型信息；所述日志查询模块设置为接收查询请求，以基于所述查询请求查询所述集群得到所述大语言模型的调用记录信息。

WO 2026/188656 A1

[见续页]

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, UY, 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))。