

(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 2024/198355 A1

(43) 国际公布日
2024 年 10 月 3 日 (03.10.2024)

(51) 国际专利分类号:

G06F 16/332 (2019.01) **G06F 16/338** (2019.01)
G06F 16/33 (2019.01) **G06F 21/60** (2013.01)

(21) 国际申请号: PCT/CN2023/129237

(22) 国际申请日: 2023 年 11 月 2 日 (02.11.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310328971.8 2023 年 3 月 30 日 (30.03.2023) CN

(72) 发明人: 及

(71) 申请人: 杨子言 (YANG, Ziyang) [CN/CN]; 中国福建省福州市鼓楼区八一七北路 189 号安泰中心 A 区 4 座 707 单元, Fujian 350000 (CN)。

(74) 代理人: 福州市京华专利代理事务所 (普通合伙) (FU ZHOU JING HUA PATENT AGENCY (GP)); 中国福建省福州市鼓楼区软件大道 89 号福州软件园 F 区 2 号楼大鹏智慧中心楼层标号 7 层 701 单元-1, Fujian 350003 (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。

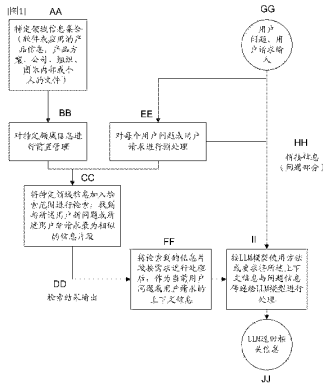
(54) Title: HUMAN-AI INTERACTION METHOD AND SYSTEM BASED ON LLM MODEL

(54) 发明名称: 一种基于 LLM 模型的人智交互方法和系统

(57) Abstract: Provided in the present invention are a method and a system. The method comprises: performing lead management on specific domain information, the specific domain information being the information which is not involved in training data of a LLM model; when user questions or user requests are received, preprocessing each user question or user request, the user question or user request referring to the user question or user request without preceding information; adding the specific domain information into a retrieval range for retrieval, so as to find an information fragment most similar to the user question or user request; processing the retrieved information fragment according to requirements, so as to use same as contextual information for the current user question or user request; and, according to a use method or requirement of a LLM model interface, transmitting the contextual information and the question information to the LLM model for processing. Therefore, the present invention solves the problems of limitation on long contexts, model data timeliness and limitation, cold start, etc.

(57) 摘要: 本发明提供一种方法和系统, 方法包括: 对特定领域信息进行前置管理; 该特定领域信息为 LLM 模型训练数据未涉及的信息; 当接收到用户问题或用户请求时, 对每个用户问题或用户请求进行预处理; 用户问题或用户请求是指没有上文信息的信息; 将特定领域信息加入检索范围进行检索, 找到与用户问题或用户请求最为相似的信息片段; 将检索到的信息片段按需求进行处理后, 作为当前用户问题或用户请求的上下文信息; 按 LLM 模型接口的使用方法或要求将上下文信息与问题信息传递给 LLM 模型进行处理。从而解决长上下文限制、模型数据时效性、局限性、冷启动等问题。

WO 2024/198355 A1



AA Specific domain information set (production information and product scheme of software or applications, and internal or personal files of corporations, organizations and groups)
BB Perform lead management on the specific domain information
CC Add the specific domain information into a retrieval range for retrieval, so as to find an information fragment most similar to a new user question or new user request
DD Output a retrieval result
EE Pre-process each user question or user request
FF Process the retrieved information fragment according to requirements, so as to use same as contextual information for the current user question or user request
HH Splice information (the question part)
II According to a use method or requirement of a LLM model, transmit the contextual information or the question information to the LLM model for processing
JJ The LLM returns related information

[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))。