



(12) 发明专利申请

(10) 申请公布号 CN 122679143 A

(43) 申请公布日 2026. 09. 01

(21) 申请号 202610775904.4

(22) 申请日 2026.06.01

(71) 申请人 深圳市形意智能科技有限公司

地址 518000 广东省深圳市宝安区新安街
道兴东社区69区洪浪北二路30号信义
领御研发中心1栋905

(72) 发明人 梁勇 刘新云 吴文辉

(74) 专利代理机构 深圳市道勤知酷知识产权代
理事务所(普通合伙) 44439

专利代理师 石慧

(51) Int. Cl.

H04L 67/10 (2022.01)

G06N 5/04 (2023.01)

H04L 9/40 (2022.01)

权利要求书3页 说明书6页 附图1页

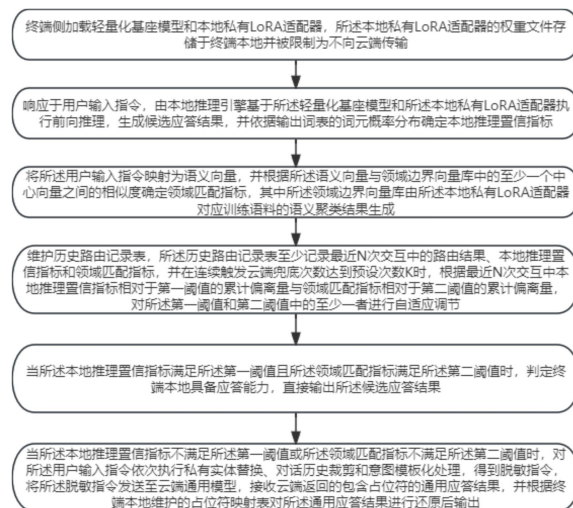
(54) 发明名称

一种基于本地LoRA优先推理的云边协同交互方法

(57) 摘要

本发明涉及人工智能与边云协同技术领域，公开了一种基于本地LoRA优先推理的云边协同交互方法、终端设备、系统及存储介质。该方法包括：终端侧加载轻量化基座模型和本地私有LoRA适配器；响应于用户输入指令，先由本地推理引擎生成候选应答结果，并基于输出词元概率分布确定本地推理置信指标；同时将用户输入指令编码为语义向量，并基于该语义向量与领域边界向量库中的中心向量之间的相似度确定领域匹配指标；根据历史路由记录对本地应答判定阈值进行自适应调节；当本地推理置信指标和领域匹配指标同时满足对应阈值时，直接输出本地候选应答结果；否则，对用户输入指令执行私有实体替换、对话历史裁剪和意图模板化处理后发送至云端通用模型，接收云端返回的包含占位符的通用应答结果，并在终端本地完成占位符还原后输出。本发明能够在降低隐私数据上传风险的同

时，提高本地命中率、交互响应速度和云端算力利用效率。



1. 一种基于本地LoRA优先推理的云边协同交互方法,其特征在于,应用于终端设备,包括:

终端侧加载轻量化基座模型和本地私有LoRA适配器,所述本地私有LoRA适配器的权重文件存储于终端本地并被限制为不向云端传输;

响应于用户输入指令,由本地推理引擎基于所述轻量化基座模型和所述本地私有LoRA适配器执行前向推理,生成候选应答结果,并依据输出词表的词元概率分布确定本地推理置信指标;

将所述用户输入指令映射为语义向量,并根据所述语义向量与领域边界向量库中的至少一个中心向量之间的相似度确定领域匹配指标,其中所述领域边界向量库由所述本地私有LoRA适配器对应训练语料的语义聚类结果生成;

终端维护历史路由记录表,所述历史路由记录表至少记录最近N次交互中的路由结果、本地推理置信指标和领域匹配指标,例如,N可以取50,即记录最近50次交互数据;N也可根据设备存储容量调整,但建议不少于20,并在连续触发云端兜底次数达到预设次数K时,根据最近N次交互中本地推理置信指标相对于第一阈值的累计偏离量与领域匹配指标相对于第二阈值的累计偏离量,对所述第一阈值和第二阈值中的至少一者进行自适应调节;

当所述本地推理置信指标满足所述第一阈值且所述领域匹配指标满足所述第二阈值时,判定终端本地具备应答能力,直接输出所述候选应答结果;

当所述本地推理置信指标不满足所述第一阈值或所述领域匹配指标不满足所述第二阈值时,对所述用户输入指令依次执行私有实体替换、对话历史裁剪和意图模板化处理,得到脱敏指令,将所述脱敏指令发送至云端通用模型,接收云端返回的包含占位符的通用应答结果,并根据终端本地维护的占位符映射表对所述通用应答结果进行还原后输出。

2. 根据权利要求1所述的云边协同交互方法,其特征在于,所述终端侧加载轻量化基座模型和本地私有LoRA适配器,包括:

读取存储于终端本地安全存储区域中的加密量化LoRA适配器权重文件;

在运行时对所述LoRA适配器权重文件执行解密和反量化处理,并加载至终端内存;

加载参数量不大于30亿的轻量化基座模型;

将所述LoRA适配器对应的低秩增量参数加载至所述轻量化基座模型的预定网络层,以形成终端本地推理单元。

3. 根据权利要求1所述的云边协同交互方法,其特征在于,所述依据输出词表的词元概率分布确定本地推理置信指标,包括:

获取本地推理引擎对当前轮次输出词表中各词元的预测概率;

根据所述预测概率的平方和确定概率集中度H;

依据词表大小V对所述概率集中度H进行归一化处理,得到归一化集中度C,所述归一化集中度C作为所述本地推理置信指标。

4. 根据权利要求1所述的云边协同交互方法,其特征在于,所述根据所述语义向量与领域边界向量库中的至少一个中心向量之间的相似度确定领域匹配指标,包括:

采用终端本地部署的语义编码模型将所述用户输入指令映射为语义向量;

读取预存于终端本地的领域边界向量库,所述领域边界向量库中的中心向量由所述本地私有LoRA适配器训练样本的语义向量经过聚类获得;

分别计算所述语义向量与各所述中心向量之间的余弦相似度；

将所述余弦相似度中的最大值作为所述领域匹配指标。

5. 根据权利要求1所述的云边协同交互方法,其特征在于,所述对所述用户输入指令依次执行私有实体替换、对话历史裁剪和意图模板化处理,包括:

对所述用户输入指令进行实体识别,识别其中的私有实体,并使用同类型占位符替换所述私有实体,生成仅保存在终端本地的占位符映射表;

删除当前轮次之外的本地对话历史,仅保留当前轮次指令文本;

根据预置的意图泛化规则库对保留后的指令文本进行模板化重写,以获得可供云端通用模型处理的脱敏指令。

6. 根据权利要求1所述的云边协同交互方法,其特征在于,所述根据最近N次交互中本地推理置信指标相对于第一阈值的累计偏离量与领域匹配指标相对于第二阈值的累计偏离量,对所述第一阈值和第二阈值中的至少一者进行自适应调节,包括:

当最近N次交互中本地推理置信指标低于第一阈值的累计偏离量大于领域匹配指标低于第二阈值的累计偏离量时,下调第一阈值;

当最近N次交互中本地推理置信指标低于第一阈值的累计偏离量小于领域匹配指标低于第二阈值的累计偏离量时,下调第二阈值;

当所述两个累计偏离量相等时,同步下调第一阈值和第二阈值;

所述第一阈值和第二阈值均被限制在各自预设浮动区间内。

7. 根据权利要求1所述的云边协同交互方法,其特征在于,还包括异常处理步骤,所述异常处理步骤包括:

监测终端本地推理引擎的显存或内存占用状态,在占用状态达到预设资源阈值时,卸载所述本地私有LoRA适配器的至少部分权重并切换至云端兜底通道;

监测终端与云端之间的网络连接状态,在网络超时或连接失败时,启用终端本地缓存应答机制,返回与当前用户输入指令语义相近的历史本地应答结果。

8. 根据权利要求1所述的云边协同交互方法,其特征在于,所述将所述脱敏指令发送至云端通用模型,包括:

在云端侧建立与脱敏指令关联的指令缓存层;

当检测到与当前脱敏指令对应的历史缓存记录时,直接返回与所述历史缓存记录对应的通用应答结果;

当未检测到与当前脱敏指令对应的历史缓存记录时,由所述云端通用模型对所述脱敏指令执行推理,并将生成的通用应答结果写入所述指令缓存层。

9. 根据权利要求1所述的云边协同交互方法,其特征在于,还包括本地离线增量更新步骤,所述本地离线增量更新步骤包括:

在终端处于离线状态时,收集用户新产生的交互数据,所述交互数据包括用户输入指令和与之对应的有效反馈应答;

基于所述交互数据对本地私有LoRA适配器执行本地微调,并冻结所述轻量化基座模型的参数;

将更新后的LoRA适配器权重文件加密存储于终端本地,并同步更新所述领域边界向量库。

10. 根据权利要求1所述的云边协同交互方法, 其特征在于, 所述根据终端本地维护的占位符映射表对所述通用应答结果进行还原后输出, 包括:

接收云端返回的包含占位符的通用应答结果;

读取终端本地维护的占位符映射表, 将所述通用应答结果中的占位符替换为对应私有实体;

将还原后的应答结果与终端本地保留的当前会话语境进行拼接后输出给用户。

11. 一种终端设备, 其特征在于, 包括处理器、存储器以及存储在所述存储器中并可由所述处理器执行的计算机程序, 所述计算机程序在被所述处理器执行时, 使所述终端设备执行权利要求1至10任一项所述的方法。

12. 一种云边协同交互系统, 其特征在于, 包括终端设备和云端服务;

所述终端设备用于执行权利要求1至10任一项所述的方法中的终端侧步骤;

所述云端服务用于接收所述终端设备发送的脱敏指令, 对所述脱敏指令执行通用推理或缓存命中处理, 并向所述终端设备返回包含占位符的通用应答结果。

13. 一种计算机可读存储介质, 其上存储有计算机程序, 其特征在于, 所述计算机程序在被处理器执行时实现权利要求1至10任一项所述的方法。

一种基于本地LoRA优先推理的云边协同交互方法

技术领域

[0001] 本发明涉及人工智能、边缘计算和隐私保护交互技术领域,尤其涉及一种基于本地LoRA优先推理的云边协同交互方法、终端设备、系统及存储介质。

背景技术

[0002] 随着大模型交互能力逐步进入终端设备,现有交互方案通常在纯云端推理和纯本地推理之间进行取舍。纯云端推理能够利用较强的通用知识能力完成复杂问答,但需要持续上传用户输入指令,容易带来网络时延、云端资源拥塞以及敏感数据外传风险。纯本地推理虽然具备离线响应快和隐私数据可控的优势,但受限于终端硬件资源和本地模型容量,对陌生领域问题、长尾问题和超出训练域的问题往往缺乏稳定的应答能力。

[0003] 现有部分云边协同方案尝试按照固定规则对本地处理和云端处理进行切分,但此类方案通常存在以下问题:其一,路由条件多依赖单一阈值或静态规则,无法同时反映本地模型的即时输出确定性和当前指令与本地适配领域之间的匹配程度;其二,云端调用前缺少面向大模型交互场景的分层脱敏机制,导致用户隐私内容仍可能随指令上传;其三,阈值长期固定时容易出现本地可处理问题被过度上传、云端资源被不必要占用的情况;其四,缺乏与本地LoRA适配器相匹配的离线更新和异常降级手段,使得系统在个性化适配和复杂运行环境下的稳定性不足。

[0004] 因此,有必要提出一种能够同时兼顾本地优先响应、隐私保护、云端兜底和路由自适应能力的云边协同交互方案。

发明内容

[0005] 本发明的目的在于提供一种基于本地LoRA优先推理的云边协同交互方法,以解决上述背景技术中提到的问题。

[0006] 本发明的目的在于提供一种基于本地LoRA优先推理的云边协同交互方法、终端设备、系统及存储介质,以降低隐私数据上传风险,减少不必要的云端调用,并提升终端设备在专属场景中的交互效率和稳定性。

[0007] 为实现上述目的,本发明采用如下技术方案:

一种基于本地LoRA优先推理的云边协同交互方法,应用于终端设备,包括:终端侧加载轻量化基座模型和本地私有LoRA适配器,所述本地私有LoRA适配器的权重文件存储于终端本地并被限制为不向云端传输;响应于用户输入指令,由本地推理引擎基于所述轻量化基座模型和所述本地私有LoRA适配器执行前向推理,生成候选应答结果,并依据输出词表的词元概率分布确定本地推理置信指标;将所述用户输入指令映射为语义向量,并根据所述语义向量与领域边界向量库中的至少一个中心向量之间的相似度确定领域匹配指标;维护历史路由记录表,并在连续触发云端兜底次数达到预设次数K时,根据最近N次交互中本地推理置信指标相对于第一阈值的累计偏离量与领域匹配指标相对于第二阈值的累计偏离量,对所述第一阈值和第二阈值中的至少一者进行自适应调节;当所述本地推理置

信指标满足所述第一阈值且所述领域匹配指标满足所述第二阈值时,直接输出所述候选应答结果;当所述本地推理置信指标不满足所述第一阈值或所述领域匹配指标不满足所述第二阈值时,对所述用户输入指令依次执行私有实体替换、对话历史裁剪和意图模板化处理,得到脱敏指令,将所述脱敏指令发送至云端通用模型,接收云端返回的包含占位符的通用应答结果,并根据终端本地维护的占位符映射表对所述通用应答结果进行还原后输出。

[0008] 优选地,所述本地推理置信指标采用对输出词元概率分布进行平方和计算并归一化得到的集中度指标。与仅依据最大概率词元或单一置信度阈值的路由方式相比,上述指标能够更稳定地反映本地模型在当前轮次的输出确定性。

[0009] 优选地,所述领域匹配指标采用用户输入指令的语义向量与领域边界向量库中的中心向量之间的最大余弦相似度。通过将即时推理确定性与训练域匹配程度联合用于路由判定,能够降低本地模型在陌生领域下的误答风险。

[0010] 优选地,所述脱敏指令由私有实体替换、对话历史裁剪和意图模板化处理逐级生成。该处理流程使云端接收到的文本保留任务意图而弱化个人身份、上下文细节和专属业务标识,有利于减少隐私暴露和缓存冲突。

[0011] 本公开的实施例提供的技术方案至少带来以下有益效果:

通过在终端本地优先执行LoRA适配后的推理,并将LoRA权重限制在本地存储和加载范围内,降低了专属适配参数外传风险。

[0012] 通过将本地推理置信指标与领域匹配指标联合用于路由判定,能够同时考虑本地模型对当前问题“是否有把握回答”和“是否属于擅长领域”,减少无效本地输出与不必要云端上传。

[0013] 通过在云端调用前执行私有实体替换、对话历史裁剪和意图模板化处理,可在保留核心任务意图的同时减小隐私数据暴露面。

[0014] 通过引入基于历史路由记录的阈值自适应调节机制,可在长期运行中提高本地命中率并缓解固定规则带来的资源浪费。

[0015] 通过异常处理、本地缓存应答和离线增量更新机制,可提高终端设备在网络波动、资源紧张和长期个性化使用场景下的连续可用性。

附图说明

[0016] 图1为本发明的方法框架结构示意图。

具体实施方式

[0017] 下面将结合实施例对本发明的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0018] 实施例一

本实施例提供一种基于本地LoRA优先推理的云边协同交互方法,应用于终端设备。如图1所示,本实施例的方法包括以下核心步骤。所述终端设备可以为智能眼镜、移动终端、工业手持设备或其他具备本地模型推理能力的边缘设备。

[0019] 在本实施例中,终端侧预先部署轻量化基座模型和本地私有LoRA适配器。所述轻量化基座模型用于提供通用语义理解和生成能力,所述本地私有LoRA适配器用于针对预定业务场景提供专属领域增量能力。LoRA适配器权重文件存储于终端本地安全存储区域,可采用加密和量化存储方式。终端在运行时对LoRA适配器权重文件进行解密和反量化处理,并将相应低秩增量参数加载到轻量化基座模型的预定网络层中,以形成用于本地推理的推理单元。

[0020] 用户发起交互后,终端接收用户输入指令。终端调用本地推理引擎基于所述轻量化基座模型和LoRA适配器执行前向推理,得到候选应答结果。与此同时,终端从输出词表中提取各词元的预测概率,并根据该概率分布计算本地推理置信指标。

[0021] 在一个可选实现中,本地推理置信指标通过以下方式计算:设输出词表大小为 V ,输出词表中第 i 个词元的预测概率为 p_i ,则概率集中度 H 满足 $H = \sum (p_i)^2$;在此基础上,根据词表大小 V 对 H 执行归一化处理,得到归一化集中度 C ,其中 $C = (H - 1/V) / (1 - 1/V)$ 。由于 C 的数值区间稳定在0至1之间,因此可将 C 作为不同终端模型之间更易统一的本地推理置信指标。第一阈值即针对所述归一化集中度 C 设定。

[0022] 除即时推理结果之外,本实施例还对用户输入指令与本地LoRA适配器训练域之间的匹配程度进行判断。为此,终端调用本地语义编码模型将用户输入指令映射为语义向量。所述语义编码模型为独立于轻量化基座模型的轻量级语义表征模型(例如基于Transformer的12层768维模型,如TinyBERT或MiniLM)。该模型预先在所述本地私有LoRA适配器的训练语料上进行微调,使其输出向量空间能够有效区分不同领域的语义边界。微调时采用余弦相似度作为损失函数,教师模型可使用通用句向量模型(如text2vec-large)。微调后的模型转换为ONNX或TensorRT Lite格式,存储于终端本地,单次编码推理耗时控制在50毫秒以内。终端设有独立的语义编码推理引擎,所有编码运算均在本地完成,不依赖云端。与之对应,终端本地还维护一个领域边界向量库。所述领域边界向量库中的中心向量由LoRA适配器训练样本的语义向量聚类得到,每一个中心向量代表LoRA适配器擅长处理的一类领域分布。

[0023] 在一个具体实现中,聚类采用K-means算法,预设聚类类别数 $K=8$,以余弦距离作为相似度度量,收敛阈值设为0.0001,最大迭代次数为100次。每个簇的中心向量取簇内所有样本向量的算术平均值。所述领域边界向量库以独立文件形式存储于终端本地非易失性存储区域,仅本地推理进程可读取。

[0024] 终端分别计算当前用户输入指令的语义向量与各中心向量之间的余弦相似度,并取最大值作为领域匹配指标。第二阈值即针对所述领域匹配指标设定。这样,系统在路由前同时获得了“本地模型对当前输出的确定性”和“当前问题是否位于本地适配训练域内”两个指标。

[0025] 为了避免固定阈值长期不适配,终端还维护历史路由记录表。所述历史路由记录表至少记录最近 N 次交互中的路由结果、本地推理置信指标和领域匹配指标。当系统连续触发云端兜底的次数达到预设次数 K 时,终端启动阈值自适应调节。例如, K 可以取10,即连续10次触发云端兜底后启动调节; K 也可根据设备交互频率动态设置,但建议取值在5~20之间。

[0026] 具体地,终端计算最近 N 次交互中本地推理置信指标低于第一阈值的累计偏离量

D1,以及领域匹配指标低于第二阈值的累计偏离量D2。所述累计偏离量定义为:对于最近N次交互中所有满足‘本地推理置信指标<第一阈值’的交互轮次,将各轮次中第一阈值与本地推理置信指标的差值进行累加,即 $D1 = \sum_{i \in S} (T1 - C_i)$,其中S为最近N次交互中本地推理置信指标低于第一阈值T1的交互轮次集合, C_i 为第i轮的本地推理置信指标。

[0027] 类似地,D2为第二阈值与各轮领域匹配指标差值的累加和。通过比较D1与D2的大小,确定近期阻碍本地命中的主导因素。

[0028] 若D1大于D2,则说明近期阻碍本地命中的主要因素为本地推理置信指标,于是优先下调第一阈值;若D2大于D1,则优先下调第二阈值;若D1与D2相同,则同步下调两个阈值。阈值调节后仍限定在各自预设浮动区间内,以防止系统出现过度放宽而导致误答率升高。

[0029] 所述第一阈值和第二阈值的初始值可通过离线测试确定:在部署前,利用所述本地私有LoRA适配器训练语料的留出验证集,以本地应答准确率不低于95%且云端调用率不高于30%为目标,采用网格搜索法选取最优初始阈值。作为示例,第一阈值初始值设为0.85,浮动区间为[0.70,0.95];第二阈值初始值设为0.72,浮动区间为[0.60,0.85]。不同终端设备可根据实际算力和场景进行调整。

[0030] 完成上述指标计算后,终端执行路由判定:当本地推理置信指标满足第一阈值且领域匹配指标满足第二阈值时,直接采用本地候选应答结果作为最终输出。由于该应答结果完全在终端本地生成,因此无需上传用户输入指令,可优先满足高时效和高隐私要求的交互场景。

[0031] 若本地推理置信指标或领域匹配指标未达到对应阈值,则终端进入云端兜底流程。在进入云端之前,终端并不直接上传原始用户输入指令,而是对原始用户输入指令执行多级脱敏。

[0032] 首先,终端对用户输入指令执行实体识别,识别用户姓名、联系方式、账号标识、地理位置、业务专名等私有实体,并以同类型占位符替换所述私有实体。终端同时生成占位符映射表,所述占位符映射表仅保存在终端本地,用于在云端返回结果后恢复原始实体。

[0033] 再者,终端对当前会话的本地历史上下文进行裁剪,仅保留当前轮次直接相关的指令文本,而不将前序完整会话历史上传至云端。该处理既可减少传输数据量,也可降低历史文本中隐私信息被带出的概率。

[0034] 最后,终端根据预置的意图泛化规则库对当前轮次指令进行模板化重写。所述意图泛化规则库可以将时间、地点、动作表达和专有场景描述抽象为标准化操作模板。

[0035] 规则库基于所述LoRA适配器训练样本中的高频用户指令,通过共现词统计与人工校验构建,包含动词同义词簇(如‘查一下/看看/核对’映射为‘查询’)、时间泛化标记(如‘明天/下周/昨天’映射为‘某时间’)、地点泛化标记(如‘北京分公司/总部’映射为‘某地点’)以及专有名词占位规则。模板按意图类别分为查询、办理、修改、删除四类,每类模板由‘主语-谓语-宾语-条件’四级槽位构成。

[0036] 例如,原始指令‘帮我查一下张三明天的考勤’经模板化重写后得到‘查询某员工在某时间的考勤信息’;‘把李四的手机号改成13800000000’重写为‘修改某员工的联系方式为某号码’。经此处理,脱敏指令保留了核心任务意图,但消除了具体实体和个性化表述。

[0037] 经此处理后,终端获得既保留核心任务意图又去除了多数专属识别特征的脱敏指令。

[0038] 终端将所述脱敏指令发送至云端服务。云端服务部署有通用大模型推理单元,并可选地设置脱敏指令缓存层。云端缓存层采用键-值存储结构,键为脱敏指令的完整文本字符串,值为对应的通用应答结果。命中判断采用完全匹配(字符串精确相等),因为脱敏指令已经过模板化处理,相同意图的用户指令经脱敏后文本形式一致。未命中缓存的请求才触发云端通用模型推理,推理结果与脱敏指令绑定后存入缓存,缓存有效期设为24小时,从而减少重复推理开销;对于未命中缓存的脱敏指令,云端服务调用通用模型完成推理生成,并将生成结果写入缓存。

[0039] 云端服务返回的应答结果包含与终端本地占位符规则一致的占位符。终端接收该通用应答结果后,读取终端本地保存的占位符映射表,将所述通用应答结果中的占位符替换为对应私有实体,并与当前会话语境进行拼接和整理后输出给用户。由于实体恢复过程完全在终端本地完成,因此云端无需接触真实私有实体。

[0040] 通过上述流程,本实施例实现了本地优先推理、双指标联合路由、多级脱敏、本地占位符恢复和历史阈值自适应的完整闭环,有效平衡了隐私保护、响应速度和云端资源利用效率。

[0041] 实施例二

在实施例一的基础上,本实施例进一步提供异常处理机制。

[0042] 当终端监测到本地推理引擎的显存或内存占用达到预设资源阈值时,系统可卸载本地私有LoRA适配器的至少部分权重,并切换至云端兜底通道,以避免终端因资源耗尽而发生卡顿或崩溃。作为示例,所述预设资源阈值可设为总显存的90%或总内存的85%,当占用达到该阈值时触发卸载操作。不同终端设备可根据硬件配置调整该阈值。

[0043] 当终端监测到与云端之间的网络连接异常时,系统可启用终端本地缓存应答机制。所述本地缓存应答机制可基于历史成功应答记录建立语义索引,对当前用户输入指令与历史语义向量进行近邻匹配。具体地,采用余弦相似度计算当前指令语义向量与历史应答对应指令语义向量之间的相似度,当最高相似度大于等于0.75时,返回对应的历史应答文本(仅做占位符还原处理,不做语义改写),否则不返回。该机制保证了在弱网或断网环境下仍能提供基础交互能力。

[0044] 实施例三

在实施例一的基础上,本实施例进一步提供本地离线增量更新机制。

[0045] 当终端处于离线状态时,终端可以收集后续交互过程中产生的新增样本,包括用户输入指令和与之对应的有效反馈应答。其中,‘有效反馈应答’通过以下至少一种规则自动判定:(1)用户在收到应答后点击了‘有用’或‘满意’按钮;

(2)应答展示后用户未在5秒内发起新的提问或修改指令,且未点击‘不满意’;

(3)用户复制、分享或收藏了该应答内容;

(4)本轮应答为云端兜底返回且本地推理置信指标高于0.90,同时用户未给出负面反馈。仅满足上述任一条件的样本才被纳入微调数据集。终端利用所述新增样本对本地私有LoRA适配器执行本地微调,并在微调期间冻结轻量化基座模型的参数,以将训练影响限制在LoRA增量参数范围内。更新完成后,终端对新的LoRA适配器权重文件进行加密存储,并基于更新后的训练样本重新生成或更新领域边界向量库中的中心向量。这样,终端能够在不依赖云端训练的前提下持续积累适配自身使用场景的个性化能力。

[0046] 实施例四

本实施例提供一种终端设备。所述终端设备包括处理器、存储器以及存储于所述存储器中的计算机程序。所述计算机程序在被所述处理器执行时,使所述终端设备执行实施例一至实施例三任一实施例所述的步骤。

[0047] 实施例五

本实施例提供一种云边协同交互系统。所述系统包括终端设备和云端服务。终端设备用于执行本地优先推理、双指标路由判定、历史阈值调节、脱敏处理和本地占位符恢复;云端服务用于接收终端发送的脱敏指令,对所述脱敏指令执行通用推理或缓存命中处理,并将结果返回终端。

[0048] 实施例六

本实施例提供一种计算机可读存储介质,其上存储有计算机程序。所述计算机程序在被处理器执行时,实现实施例一至实施例三任一实施例所述的方法。

[0049] 应当理解,在不脱离本发明核心构思的前提下,上述各实施例中涉及的模型规模、阈值取值、聚类算法、缓存策略、实体类别、模板化规则和终端设备类型均可根据实际应用需求进行调整。本领域技术人员基于本发明公开内容作出的等同替换或变形,均应落入本发明的保护范围。

[0050] 以上,仅为本申请的具体实施方式,但本申请的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本申请揭露的技术范围内,可轻易想到变化或替换,都应涵盖在本申请的保护范围之内。因此,本申请的保护范围应以权利要求的保护范围为准。

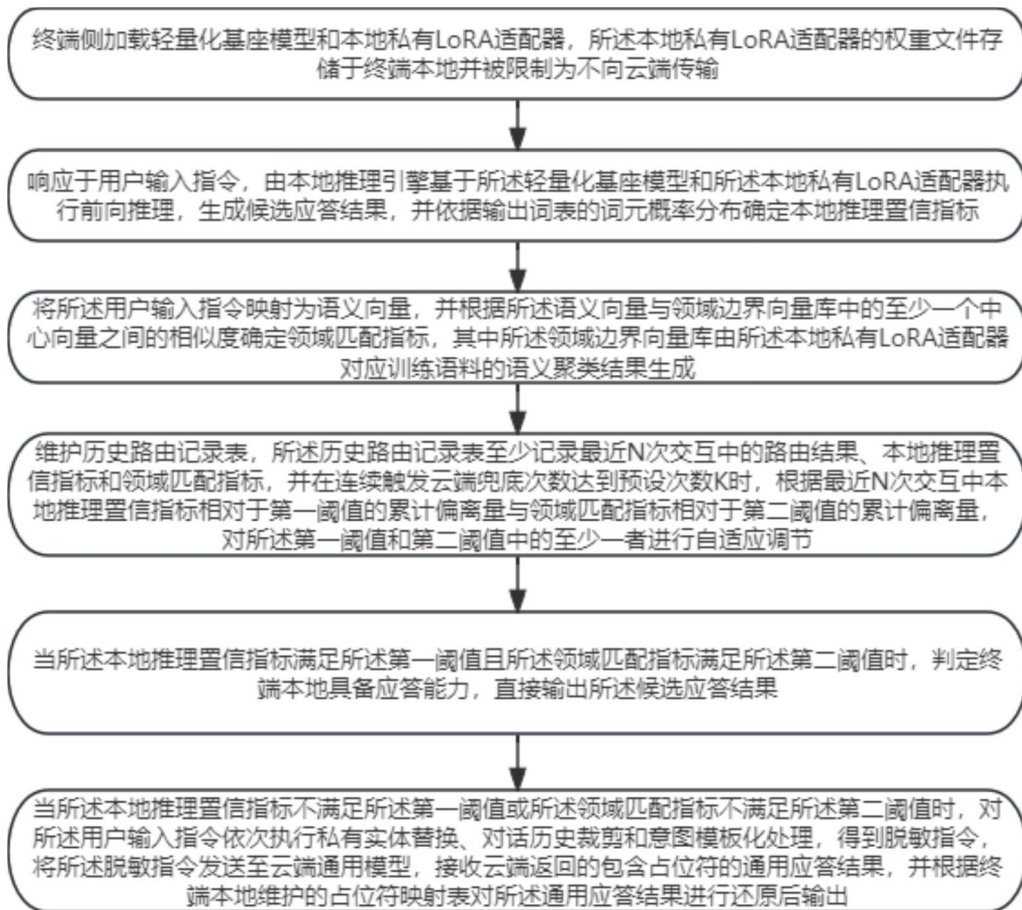


图1