

说明书

发明名称: 一种基于LLM模型的人智交互方法和系统

技术领域

[0001] 本发明涉及人智交互技术领域，特别涉及一种基于LLM模型的人智交互方法和系统。

背景技术

[0002] LLM (Large Language Model, 大型语言模型) 是基于大型标注数据训练获得的一种语言交互模型，并提供新型文字化的交互接口，供产品接入，让用户可以以口语化的方式进行交互问答；可以被广泛应用于智能客服机器人、在线文档（如文档、表格、邮件、会议纪要、日程待办、日程待办等办公文档）编辑助手和软件智能助手等人智交互产品中。

[0003] 用户在使用上述人智交互产品时会发现，在首次问询时用户需要补充提问背景，并以类似文本模板的形式提问（即使用{历史数据对话背景}+{查询输入}的模板进行交互），并在后续对话中，不断补充前期对话内容观点作为次轮或后续对话背景进行提问拓展，以获得更好的反馈结果；例如：

[0004] i. 用户与智能客服机器人进行人智对话时，首轮问询需要补充问询背景，即以类似下属文本模板进行提问：“你是【某某业务】的专家/客服人员，请针对下述问题【问题内容】，提供可能解决的办法”；次轮、后续对话需补充观点进行提问，“作为【某某业务】专家/客服人员，【某问题】的【某解决方案】进一步的分步说明”；

[0005] ii. 用户与在线文档编辑助手进行人智对话时，首轮问询需要补充问询背景，即以类似下属文本模板进行提问：“你是一名【职业/岗位】，我发给你一封电子邮件，【邮件内容】，你要回复这封邮件”；次轮、后续对话需补充观点进行提问，例如：“作为【某某领域】专家，针对邮件中【某观点】补充纲要”；

[0006] iii. 用户与在线文档编辑助手进行人智对话时，首轮问询需要补充问询背景，即以类似下属文本模板进行提问，“你是【行业】的【职位/岗位】，现在有一个电子表格，【字段1、字段2、字段3...】，需要对该表格【某些字段】进行可

可视化展示，输出【图例】”；次轮、后续对话需补充观点进行提问，“输出的【图例】有【某些问题】，需进行调整”。

[0007] 同时，为实现便捷、多轮、同一上下文交流，现有的人智交互产品存在如下缺陷的至少之一：

[0008] i. LLM对于接口token（可在模型中处理的最小文本单元）数量有限制，以chatgpt为例，当前最大token为4096，LLM对token大小的限制基于模型性能与服务的均衡考量；

[0009] ii. LLM为获得更好的数据，需要用户补充上下文信息；

[0010] iii. LLM对于上下文记忆最大字数限制，累计对话超出限制，最早的对话会被移除；

[0011] iv. 用户为绕开上述限制，需要用户自行对交互上下文信息进行合理管理，记录上下文、总分拆解分点交互、或者阶段性地总结原始对话，重新作为上下文传入，但会导致要压缩信息，对信息有损。

[0012] 究其原因，是由于现有的LLM模型其本身的特性，存在如下缺陷：

[0013] （1）LLM模型无状态特性

[0014] LLM模型特性是无状态的，意味着每个传入查询的处理都是独立于其他交互，无状态交互需要有记忆管理，例如使用{对话背景}+{查询输入}的模板进行交互；

[0015] （2）长上下文限制

[0016] LLM模型自身特点是需要对token进行处理，模型需在token大小（词汇表）和计算、存储资源做平衡，所以token数量有限，对长上下文不友好，另一方面，为突破限制，结合使用阶段性总结对话作为上下文，导致要压缩信息，对信息有损；

[0017] （3）具有模型数据时效性

[0018] LLM模型是基于数据预训练，反馈的信息具有时效性，即训练数据是在某些特定年份前的信息，训练数据后续产生的信息未参与训练，导致模型反馈的信息不准确；

[0019] （4）模型数据具有局限性

[0020] 当使用LLM模型询问特定领域信息（即LLM模型训练数据未涉及的信息，如公司内部流程机制、规范、手册、产品介绍、开发内容等私域信息），LLM无法给出准确答案；例如当用户有对表格文档处理的需求时，当前LLM基于训练数据(数据可能包含软件编辑使用信息)，反馈内容局限在文档编辑，无法对具体文档内容进行处理；

[0021] （5）存在对话轮次长、冷启动的问题

[0022] 对于特定领域信息（即LLM模型训练数据未涉及的信息，如公司内部流程机制、规范、手册、产品介绍、开发内容等私域信息）的问答，为了得到更准确的反馈信息，可能需要多轮对话，对上下文进行反馈补充，才能解决冷启动问题。

发明概述

技术问题

[0023] 本发明要解决的技术问题，在于提供一种基于LLM模型的人智交互方法和系统，在LLM模型的训练数据中不涉及特定领域信息时的使用场景，能为用户自动补充相关上下文信息（也称辅助信息），将上下文信息和问题信息传递给LLM模型行使用，从而解决长上下文的限制、模型数据时效性、局限性、冷启动等问题。

技术方案

[0024] 第一方面，本发明提供了一种基于LLM模型的人智交互方法，包括下述步骤：

[0025] S1、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；

[0026] S2、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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；

[0027] S3、将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；

[0028] S4、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；

- [0029] S5、按LLM模型接口使用方法或要求将所述上下文信息与问题信息传递给LLM模型进行处理；
- [0030] 其中，所述S1和所述S2不限先后顺序。
- [0031] 第二方面，本发明提供了一种基于LLM模型的人智交互系统，包括：
- [0032] 前置管理模块，用于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；
- [0033] 问题预处理模块，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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；
- [0034] 检索模块，用于将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0035] 信息片段处理模块，用于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；
- [0036] 传递模块，用于按LLM模型接口的使用方法或要求将所述上下文信息与问题信息传递给LLM模型进行处理。

有益效果

- [0037] 本发明实施例中提供的一个或多个技术方案，至少具有如下技术效果或优点：通过特定领域信息的前置管理并加入检索范围进行检索，为LLM模型补充训练数据未涉及的信息，为用户问题或用户请求自动补充上下文信息，解决了长文本的长上下文限制、模型数据时效性、局限性、冷启动等问题，从而适于更多特定使用场景中。通过前置信息管理、问题预处理、信息检索、自动拼接的方式，易于实现LLM模型的产品化，满足例如在特殊使用场景如智能客服(包括问答机器人、营销机器人)、在线文档编辑助手、软件智能助手中的应用；基于前置信息管理，降低用户多轮补充录入上下文的情况，一定程度上避免用户手工录入上下文的麻烦，提升用户好感度，且对于收费的LLM模型而言，还可在一定程度上降低访问次数，降低产品的使用成本。

[0038] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0039] 下面参照附图结合实施例对本发明作进一步的说明。

[0040] 图1为本发明实施例中方法中的流程图；

[0041] 图2为本发明实施例中系统的结构示意图。

本发明的实施方式

[0042] 本申请实施例通过提供一种基于LLM模型的人智交互方法和系统，在LLM模型的训练数据中不包含特定领域数据时的使用场景，仍能为用户自动补充相关上下文信息，将上下文信息与问题信息传递给LLM模型行使用，从而解决长上下文的限制、模型数据时效性、局限性、冷启动等问题。

[0043] 本申请实施例中的技术方案，总体思路如下：对特定领域信息进行前置管理；对每个用户问题或用户请求进行预处理；自动在公开、半公开、私域数据或上述数据组合范围中进行检索，找到与用户问题或用户请求最为相似的信息片段；将检索到的相关信息片段直接或者再次处理或者组合的形式，作为上下文信息；将上下文信息与问题信息传递给LLM模型进行处理，从而解决长上下文的限制、模型数据时效性、局限性、冷启动等问题。

[0044] 如图1所示，本申请实施例提供一种基于LLM模型的人智交互方法，包括下述步骤：

[0045] S1、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，包括但不限于：

[0046] i. 软件、或应用的产品信息、产品方案等；

[0047] ii. 商品或产品信息、产品方案等；

[0048] iii. 基于组织(企业、团队)或个人的内部文档资料信息；

[0049] iv. 软件或产品付费才能解锁的资讯、文档信息；

- [0050] 这些信息由于处于未公开状态，因此不会做为LLM模型训练数据，从而使LLM模型在使用过程中，不具有对该些数据进行检索；
- [0051] 所述前置管理具体是：将特定领域信息进行预处理，使用非向量数据库构建数据集进行管理，或者使用向量数据库构建数据集进行管理构建数据集进行存储管理，或者使用自研或者第三方搜索技术对特定领域信息进行管理；
- [0052] S2、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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；
- [0053] S3、将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；将特定领域信息加入检索范围后，检索范围即可包括特定领域信息，但并不排除包括原有的公开数据和半公开数据；
- [0054] S4、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；
- [0055] S5、按LLM模型接口的使用方法或要求将所述上下文信息与问题信息传递给LLM模型进行处理；
- [0056] 其中，所述S1和所述S2不限先后顺序。
- [0057] 进一步的，所述S4中，所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- [0058] 所述S4中，将检索到的信息片段按需求进行处理后，还作为未来LLM接口要求的数据；
- [0059] 所述S5中，所述传递为按LLM接口使用方法或要求对信息进行处理后传递，包括先将所述上下文信息与问题信息进行组合拼接后再进行传递，或分别通过独立接口进行传递，或先进行加密后再进行传递。
- [0060] 基于同一发明构思，本申请还提供了与上述方法对应的装置，如图2所示，本申请实施例提供了一种基于LLM模型的人智交互系统，包括：
- [0061] 前置管理模块，用于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；

- [0062] 问题预处理模块，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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；
- [0063] 检索模块，用于将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0064] 信息片段处理模块，用于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；
- [0065] 传递模块，用于按LLM模型接口的使用要求或要求将所述上下文信息与问题信息传递给LLM模型进行处理。
- [0066] 其中，所述信息片段处理模块中，所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- [0067] 所述信息片段处理模块中，将检索到的信息片段按需求进行处理后，还作为未来LLM接口要求的数据；
- [0068] 所述传递包括先将所述上下文信息与问题信息进行组合拼接后再进行传递，或分别通过独立接口进行传递，或先进行加密后再进行传递。
- [0069] 实施例一
- [0070] 作为本发明的一种更为具体的实现方式，当所述前置管理为通过使用非向量数据库构建数据集进行管理时，所述人智交互方法具体包括：
- [0071] S11、首先将特定领域信息进行预处理，包括对内容分词解析，并将分词标签作为对应信息内容的元数据，使用非向量数据库构建数据集进行管理；
- [0072] S12、当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行分词解析；或者还包括本地对问题的存储管理，以适于阶段性对话；
- [0073] S13、将特定领域信息加入检索范围，通过分词权重关键词进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0074] S14、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；所述处理为直接处理和/或再次处理；所述再次处理为对检索出

的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；

[0075] S15、按LLM模型接口的使用方法或要求将所述上下文信息和问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

[0076] 实施例二

[0077] 当所述前置管理为通过使用向量数据库构建数据集进行管理时，所述人智交互方法具体包括：

[0078] S21、首先将特定领域信息进行预处理，包括对特定领域信息进行向量化处理，并使用向量数据库进行管理；

[0079] S22、当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行向量化处理；或者还包括本地对问题的存储管理，以适于阶段性对话；

[0080] S23、将特定领域信息加入检索范围，结合embedding嵌入技术进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；

[0081] S24、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；所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；

[0082] S25、按LLM模型接口的使用方法或要求将所述上下文信息和问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

[0083] 实施例三

[0084] 当所述前置管理为基于第三方前置信息管理时，所述人智交互方法具体包括：

[0085] S31、基于自研或第三方信息搜索技术对特定领域信息进行管理；

[0086] S32、当接收到用户问题或用户请求时，基于所述自研或第三方信息搜索的要求，对所述用户问题或所述用户请求进行预处理，该预处理包括无处理、分词处理、提取关键词或本地对问题的存储管理；

[0087] S33、将特定领域信息加入检索范围，使用自研或第三方信息搜索技术进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；

- [0088] S34、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；所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- [0089] S35、按LLM模型接口的使用要求或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。
- [0090] 实施例四
- [0091] 当所述前置管理为通过使用非向量数据库构建数据集进行管理时，所述人智交互系统具体包括：
- [0092] 前置管理模块，用于将特定领域信息进行预处理，包括对内容分词解析，并将分词标签作为对应信息内容的元数据，使用关系型数据或非关系型数据的格式构建数据集进行管理；
- [0093] 问题预处理模块，当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行分词解析；或者还包括本地对问题的存储管理，以适于阶段性对话；
- [0094] 检索模块，用于将特定领域信息加入检索范围，通过分词权重关键词进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0095] 信息片段处理模块，用于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；
- [0096] 传递模块，用于按LLM模型接口的使用要求或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。
- [0097] 实施例五
- [0098] 当所述前置管理为通过使用向量数据库构建数据集进行管理时，所述人智交互方法具体包括：
- [0099] 前置管理模块，用于将特定领域信息进行预处理，包括对特定领域信息进行向量化处理，并使用向量数据库进行管理；

- [0100] 问题预处理模块，当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行向量化处理；或者还包括本地对问题的存储管理，以适于阶段性对话；；
- [0101] 检索模块，用于将特定领域信息加入检索范围，结合embedding嵌入技术进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0102] 信息片段处理模块，用于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；所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- [0103] 传递模块，按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。
- [0104] 实施例六
- [0105] 当所述前置管理为基于第三方前置信息管理时，所述人智交互方法具体包括：
- [0106] 前置管理模块，用于基于自研或第三方信息搜索技术对特定领域信息进行管理；
- [0107] 问题预处理模块，当接收到用户问题或用户请求时，基于所述自研或第三方信息搜索的要求，对所述用户问题或所述用户请求进行预处理，该预处理包括无处理、分词处理、提取关键词或本地对问题的存储管理；
- [0108] 检索模块，用于将特定领域信息加入检索范围，使用自研或第三方信息搜索技术进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；
- [0109] 信息片段处理模块，用于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；所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；

- [0110] 传递模块，按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。
- [0111] 由于本发明实施例四至六所介绍的系统，为实施本发明实施例一至三的方法所采用的装置，故而基于本发明实施例一至三所介绍的方法，本领域所属人员能够了解该装置的具体结构及变形，故而在此不再赘述。凡是本发明实施例一至三的方法所采用的装置都属于本发明所欲保护的范围。
- [0112] 本申请实施例至少具有如下技术效果或优点：通过特定领域信息的前置管理并加入检索范围进行检索，为LLM模型补充训练数据未涉及的信息，为用户问题或用户请求自动补充上下文信息，解决了长文本的长上下文限制、模型数据时效性、局限性、冷启动等问题，从而适于更多特定使用场景中。通过前置信息管理、问题预处理、信息检索、自动拼接的方式，易于实现LLM模型的产品化，满足例如在特殊使用场景如智能客服(包括问答机器人、营销机器人)、在线文档编辑助手、软件智能助手中的应用；基于前置信息管理，降低用户多轮补充录入上下文的情况，一定程度上避免用户手工录入上下文的麻烦，提升用户好感度，且对于收费的LLM模型而言，还可在一定程度上降低访问次数，降低产品的使用成本。
- [0113] 虽然以上描述了本发明的具体实施方式，但是熟悉本技术领域的技术人员应当理解，我们所描述的具体的实施例只是说明性的，而不是用于对本发明的范围的限定，熟悉本领域的技术人员在依照本发明的精神所作的等效的修饰以及变化，都应当涵盖在本发明的权利要求所保护的范围内。

权利要求书

- [权利要求 1] 一种基于LLM模型的人智交互方法，其特征在于：包括下述步骤：
- S1、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；
- S2、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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；
- S3、将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；
- S4、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；
- S5、按LLM模型接口使用方法或要求将所述上下文信息与问题信息传递给LLM模型进行处理；
- 其中，所述S1和所述S2不限先后顺序。
- [权利要求 2] 根据权利要求1所述的一种基于LLM模型的人智交互方法，其特征在于：所述S4中，所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- 所述S4中，将检索到的信息片段按需求进行处理后，还作为未来LLM接口要求的数据；
- 所述S5中，所述传递包括先将所述上下文信息与问题信息进行组合拼接后再进行传递，或分别通过独立接口进行传递，或先进行加密后再进行传递。
- [权利要求 3] 根据权利要求1或2所述的一种基于LLM模型的人智交互方法，其特征在于：当所述前置管理为使用非向量数据库构建数据集进行管理时，所述人智交互方法具体包括：

S11、首先将特定领域信息进行预处理，包括对内容分词解析，并将分词标签作为对应信息内容的元数据，使用关系型数据或非关系型数据的格式构建数据集进行管理；

S12、当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行分词解析，或者还包括本地对问题的存储管理；

S13、将特定领域信息加入检索范围，通过分词权重关键词进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；

S14、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；

S15、按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

[权利要求 4]

根据权利要求1或2所述的一种基于LLM模型的人智交互方法，其特征在于：当所述前置管理为通过使用向量数据库构建数据集进行管理时，所述人智交互方法具体包括：

S21、首先将特定领域信息进行预处理，包括对特定领域信息进行向量化处理，并使用向量数据库进行管理；

S22、当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行向量化处理，或者还包括本地对问题的存储管理；

S23、将特定领域信息加入检索范围，结合embedding嵌入技术进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；

S24、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；

S25、按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

[权利要求 5]

根据权利要求1或2所述的一种基于LLM模型的人智交互方法，其特征在于：当所述前置管理为基于第三方前置信息管理时，所述人智交互方法具体包括：

S31、基于自研或第三方信息搜索技术对特定领域信息进行管理；

S32、当接收到用户问题或用户请求时，基于所述自研或第三方信息搜索的要求，对所述用户问题或所述用户请求进行预处理，该预处理包括无处理、分词处理、提取关键词或本地对问题的存储管理；

S33、将特定领域信息加入检索范围，使用自研或第三方信息搜索技术进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；

S34、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；

S35、按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

[权利要求 6]

一种基于LLM模型的人智交互系统，其特征在于：包括：

前置管理模块，用于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；

问题预处理模块，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；所述用户问题或所述用户请求是指没有上文信息的用户问题或用户请求；

检索模块，用于将特定领域信息加入检索范围进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；

信息片段处理模块，用于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；

传递模块，用于按LLM模型接口的使用方法或要求将所述上下文信息与问题信息传递给LLM模型进行处理。

- [权利要求 7] 根据权利要求6所述的一种基于LLM模型的人智交互系统，其特征在于：所述信息片段处理模块中，所述处理为直接处理和/或再次处理；所述再次处理为对检索出的信息片段进行拷贝，编辑，总结，排序，筛选，翻译，压缩、过滤、重新编码或者再次、多次重新检索、基于检索结果的代码执行的至少一种；
- 所述信息片段处理模块中，将检索到的信息片段按需求进行处理后，还作为未来LLM接口要求的数据；
- 所述传递模块中，所述传递包括先将所述上下文信息和问题信息进行组合拼接后再进行传递，或分别通过独立接口进行传递，或先进行加密后再进行传递。
- [权利要求 8] 根据权利要求6或7所述的一种基于LLM模型的人智交互系统，其特征在于：当所述前置管理为使用非向量数据库构建数据集进行管理时，所述人智交互系统具体包括：
- 前置管理模块，用于将特定领域信息进行预处理，包括对内容分词解析，并将分词标签作为对应信息内容的元数据，使用关系型数据或非关系型数据的格式构建数据集进行管理；
- 问题预处理模块，当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行分词解析，或者还包括本地对问题的存储管理；
- 检索模块，用于将特定领域信息加入检索范围，通过分词权重关键词进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；
- 信息片段处理模块，用于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；
- 传递模块，用于按LLM模型接口的使用方法或要求将所述上下文信息和问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。

- [权利要求 9] 根据权利要求6或7所述的一种基于LLM模型的人智交互系统，其特征在于：当所述前置管理为通过使用向量数据库构建数据集进行管理时，所述人智交互方法具体包括：
- 前置管理模块，用于将特定领域信息进行预处理，包括对特定领域信息进行向量化处理，并使用向量数据库进行管理；
- 问题预处理模块，当接收到用户问题或用户请求时，对所述用户问题或所述用户请求进行向量化处理，或者还包括本地对问题的存储管理；
- 检索模块，用于将特定领域信息加入检索范围，结合embedding嵌入技术进行索引，找到与所述用户问题或所述用户请求最为相似的信息片段；
- 信息片段处理模块，用于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；
- 传递模块，按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。
- [权利要求 10] 根据权利要求6或7所述的一种基于LLM模型的人智交互系统，其特征在于：当所述前置管理为基于第三方前置信息管理时，所述人智交互方法具体包括：
- 前置管理模块，用于基于自研或第三方信息搜索技术对特定领域信息进行管理；
- 问题预处理模块，当接收到用户问题或用户请求时，基于所述自研或第三方信息搜索的要求，对所述用户问题或所述用户请求进行预处理，该预处理包括无处理、分词处理、提取关键词或本地对问题的存储管理；
- 检索模块，用于将特定领域信息加入检索范围，使用自研或第三方信息搜索技术进行检索，找到与所述用户问题或所述用户请求最为相似的信息片段；

信息片段处理模块，用于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；

传递模块，按LLM模型接口的使用方法或要求将所述上下文信息与问题信息进行组合拼接传递或分别通过独立接口传递或进行加密传递给LLM模型进行处理。