说　　明　　书

测试集的构建方法及系统、模型的评估方法及系统

技术领域

本说明书涉及人工智能技术领域，尤其涉及一种测试集的构建方法及系统、模型的评估方法及系统。

背景技术

在大语言模型的应用中，检索增强生成架构（Retrieval-Augmented-Generation，简称 RAG）是一种重要的生成方式。检索增强生成系统（即 RAG 系统）是通过检索来增强大语言模型的生成能力的。检索增强生成系统是一个结合了检索和生成两部分的神经网络架构，通过检索与问题相关的信息，并将这些相关信息作为额外的信息输入到大语言模型中，可以得到对应的问题回答。因此检索增强生成系统的检索能力对于大语言模型的生成能力极其重要。而评估检索增强生成系统的一个重要方式是通过测试集进行评估。

因此，需要提供一种检索增强生成系统的测试集的构建方法，以用于对检索增强生成系统的检索能力进行评估。

背景技术部分的内容仅仅是发明人个人所知晓的信息，并不代表上述信息在本说明书申请日之前已经进入公共领域，也不代表其可以成为本说明书的现有技术。

发明内容

本说明书提供一种测试集的构建方法及系统、模型的评估方法及系统。

第一方面，本说明书提供一种检索增强生成系统的测试集的构建方法，包括：基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合；基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题；以及将所述测试问题与其对应的测试文档组合作为一组标签数据，构建测试集，所述测试集用于对所述检索增强生成系统的检索能力进行评估。

第二方面，本说明书还提供一种检索增强生成系统的评估方法，包括：获取测试集，

所述测试集是基于以下方式得到的：基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，以及将所述测试问题与其对应的测试文档组合作为一组标签数据，构建所述测试集；对于一组标签数据，将所述测试问题输入至所述检索增强生成系统的检索模型中，将所述检索模型输出的结果与所述测试问题对应的测试文档组合进行对比，得到对应的评估结果；以及基于所述测试集对应的评估结果，确定目标评估结果。

第三方面，本说明书还提供一种增强模型的训练系统，包括：至少一个存储介质，存储有至少一个指令集，用于进行增强模型的训练方法；以及至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并根据所述至少一个指令集的指示执行：基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合；基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题；以及将所述测试问题与其对应的测试文档组合作为一组标签数据，构建测试集，所述测试集用于对所述检索增强生成系统的检索能力进行评估。

第四方面，本说明书还提供一种检索增强生成系统的评估系统，包括：至少一个存储介质，存储有至少一个指令集，用于进行增强模型的训练方法；以及至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并根据所述至少一个指令集的指示执行：获取测试集，所述测试集是基于以下方式得到的：基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，以及将所述测试问题与其对应的测试文档组合作为一组标签数据，构建所述测试集；对于一组标签数据，将所述测试问题输入至所述检索增强生成系统的检索模型中，将所述检索模型输出的结果与所述测试问题对应的测试文档组合进行对比，得到对应的评估结果；以及基于所述测试集对应的评估结果，确定目标评估结果。

由以上技术方案可知，本说明书提供的测试集的构建方法及系统、模型的评估方法及系统所构建的测试集中的一组标签数据由一个测试问题和测试文档组合组成，一个测试文档组合包括相互关联的一个或多个文档块。而测试问题是基于测试文档组合生成的总结性问题。即测试问题是对测试文档组合中的问题的总结。由此方式构建的测试集在对检索增强生成系统进行评估时，不仅可以评估检索增强生成系统根据问题检索单一文

档的能力，还能评估检索增强生成系统根据总结性的问题检索关联文档的能力，从而提升对检索增强生成系统的检索能力评估的全面性。

本说明书提供的测试集的构建方法及系统、模型的评估方法及系统的其他功能将在以下说明中部分列出。本说明书提供的测试集的构建方法及系统、模型的评估方法及系统的创造性方面可以通过实践或使用下面详细示例中所述的方法、装置和组合得到充分解释。

附图说明

为了更清楚地说明本说明书实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建方法和模型的评估方法的应用场景示意图；

图 2 示出了根据本说明书的一些实施例提供的一种计算设备的硬件结构示意图；

图 3 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建方法的流程图；

图 4 示出了根据本说明书的实施例提供的一种文档块关联图谱的示意图；

图 5 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建的步骤示意图；以及

图 6 示出了根据本说明书的实施例提供的一种检索增强生成系统的评估方法的流程图。

具体实施方式

以下描述提供了本说明书的特定应用场景和要求，目的是使本领域技术人员能够制造和使用本说明书中的内容。对于本领域技术人员来说，对所公开的实施例的各种局部修改是显而易见的，并且在不脱离本说明书的精神和范围的情况下，可以将这里定义的一般原理应用于其他实施例和应用。因此，本说明书不限于所示的实施例，而是与权利要求一致的最宽范围。

这里使用的术语仅用于描述特定示例实施例的目的，而不是限制性的。比如，除非上下文另有明确说明，这里所使用的，单数形式“一”，“一个”和“该”也可以包括复数形式。当在本说明书中使用时，术语“包括”、“包含”和/或“含有”意思是指所关联的整数，步骤、操作、元素和/或组件存在，但不排除一个或多个其他特征、整数、步骤、操作、元素、组件和/或组的存在或在该系统/方法中可以添加其他特征、整数、步骤、操作、元素、组件和/或组。

考虑到以下描述，本说明书的这些特征和其他特征、以及结构的相关元件的操作和功能、以及部件的组合和制造的经济性可以得到明显提高。参考附图，所有这些形成本说明书的一部分。然而，应该清楚地理解，附图仅用于说明和描述的目的，并不旨在限制本说明书的范围。还应理解，附图未按比例绘制。

本说明书中使用的流程图示出了根据本说明书中的一些实施例的系统实现的操作。应该清楚地理解，流程图的操作可以不按顺序实现。相反，操作可以以反转顺序或同时实现。此外，可以向流程图添加一个或多个其他操作。可以从流程图中移除一个或多个操作。

为了方便描述，首先对本说明书后文中将会出现的术语进行解释。

大语言模型：通常在自然语言处理（NLP）的人工智能领域中使用，特别是在指代具有大量参数和计算资源的大型机器学习模型，大语言模型因为拥有巨大数量的参数和复杂的网络结构而得名，其具有强大的特征表示能力和特征理解能力，在处理复杂任务时能够更好地捕捉数据中的模式和规律，它们是为了能更好地理解和生成自然语言而设计和训练的。

检索增强生成架构（Retrieval-Augmented-Generation，简称 RAG）：RAG 增强检索链路是一种利用检索和生成两个子任务来提高大语言模型的能力和质量的方法。它的基本思想是，对于给定的问题或输入，首先从大规模的知识库中检索出相关信息，然后将这些相关信息作为额外的信息输入到大模型中，从而生成更准确、更具体、更有知识性的回答或输出。

实体：在模型训练中，特别是在数据建模、数据库设计、人工智能以及自然语言处理等领域中，实体（Entity）是信息和知识表示中的一个重要概念。它通常指代现实世界中的具体或抽象事物，可以是人、地点、组织、概念等。在数据库和模型设计中，实体是组织收集信息的载体。它们通过属性和关系来定义和描述现实世界中的事物。

下面对本说明书的应用场景进行介绍。

本说明书提供的测试集的构建方法及系统、模型的评估方法及系统一般应用在基于检索增强生成架构（即前述的 RAG 系统）下通过大语言模型进行检索或进行问答的场景中。主要应用场景包括：在应用大语言模型进行检索/问答之前，构建测试集，并基于测试集对检索增强生成系统的检索能力进行评估。检索增强生成系统的检索能力是指检索增强生成系统根据输入大语言模型的问题检索出与问题相关的信息的能力。检索能力可以用检索出的相关信息的准确度进行衡量。检索增强生成系统的检索能力由检索模型提供。与问题相关的信息可以作为大语言模型的提示数据。大语言模型可以基于输入的问题以及与问题相关的信息输出问题对应的答案。与问题相关的信息越准确，大语言模型得到的答案也越准确。因此，检索增强生成系统的检索能力对于大语言模型计算的准确度极其重要。

本说明书提供的测试集的构建方法及系统可以基于特定数据集构建测试集。本说明书提供的模型的评估方法及系统可以基于构建好的测试集，对检索增强生成系统针对特定数据集的检索能力进行评估。评估的结果可以反映出检索增强生成系统在特定数据集上的检索效果。检索能力越高代表检索效果越准确。反之，检索能力越低代表检索效果越不准确。当检索增强生成系统在特定数据集上的检索效果较差时，后续系统可以基于检索增强生成系统的检索能力的评估结果，对检索增强生成系统的检索模型进行调整，以提高其检索能力。

如前所述，测试集用于评估检索增强生成系统的检索能力。同时，检索增强生成系统的检索能力是指检索增强生成系统根据输入大语言模型的问题检索出与问题相关的信息的能力。也就是说，测试集中的测试数据是由问题以及与问题相关的信息组成的数据对。而与问题相关的信息来自于源数据库（即前述的特定数据集）。其中，构建测试集一般是将源数据库中的文档从数据库中取出，通过大语言模型对文档进行总结，并建立各文档与各问题之间的对应关系，以将问题和对应的文档作为一组测试数据建立测试集。

在测试集的构建场景下，一般是通过大语言模型对源数据库中的每个文档进行问题提取，并将提取的问题和原始文档作为数据对，以建立测试集。但是这样的构建方式在大语言模型的应用中，只能评估模型根据问题检索单一文档的检索能力，无法评估模型根据问题检索多个文档的检索能力。例如对于基于多个文档综合得到的总结类问题来说，

上述方法无法准确评估模型的检索能力。而在大语言模型的实际应用中，一个文档往往无法给大语言模型提供足够的信息。因此上述方法的应用具有一定的局限性，存在评估结果不够准确、不够全面的问题。

因此，本说明书提供一种检索增强生成系统的测试集的构建方法，能够对源数据库中的文档进行重新构建，将具有关联关系的文档块组成测试文档组合，抽取测试文档组合对应的测试问题，将测试问题与其对应的测试文档组合作为一组标签数据，构建测试集。本说明书提供的检索增强生成系统的测试集的构建方法不仅可以评估检索增强生成系统根据问题检索单一文档的能力，还能评估检索增强生成系统根据总结性的问题检索关联文档的能力，从而提升对检索增强生成系统的检索能力评估的全面性。所述

需要说明的是，上述有关场景只是本说明书提供的多个使用场景中的一个，本领域技术人员应当明白，当本说明书提供的测试集的构建方法及系统、模型的评估方法及系统应用于其他使用场景时，其实现方式及技术效果是类似的。

图 1 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建方法和模型的评估方法的应用场景示意图。如图 1 所示，该应用场景 100 可以包括源数据库 110、检索增强生成系统 120（以下简称 RAG 系统 120）、RAG 系统 120 的测试集的构建系统 130（以下简称构建系统 130）以及 RAG 系统 120 的评估系统 150（以下简称评估系统 150）。

源数据库 110 是前述的检索增强生成架构下的知识库。源数据库 110 中可以存储有与应用领域或应用场景相关的知识文档。所述应用领域或应用场景可以是大语言模型的应用领域或应用场景。比如，金融场景、医疗场景，等等。例如，在金融场景下，源数据库 110 中可以存储与金融领域相关的知识文档，如金融分析报告，等等。例如，在医疗场景下，源数据库 110 中可以存储与医疗领域相关的知识文档，如病理分析报告、药物治疗原理、药物用法与用量，等等。

源数据库 110 中的知识文档可以以文档块的形式存在。即对一个文档进行分块处理，以切分为多个文档块。在本说明书的实施例中，源数据库 110 中例如可以存储有大量的知识文档。本说明书中可以基于预设的切分规则对源数据库 110 中的各知识文档进行切分，以得到多个文档块（chunks）。

其中，在一些实施例中，预设的切分规则可以为：根据预设字数对源数据库 110 中的各知识文档进行切分。例如以 1000 字为预设的切分单位，将各知识文档切分为 1000

字、或少于 1000 字的多个文档块。需要说明的是一个文档块不应当包含不完整的内容。不完整的内容可以是不完整的句子。例如，当一个文档块在第 1000 个字时对应的句子不完整时，应当将此文档块中不完整的句子作为下一个文档块的内容。在一些实施例中，也可以为以预设个数的段落为预设的切分单位。比如，以 1 个段落为预设的切分单位，将各知识文档切分为多个文档块，一个文档块包含 1 个段落。或者，也可以以 2 个段落，或者 3 个段落、5 个段落为预设的切分单位，对各知识文档进行切分，以得到由预设个数段落构成的文档块。在一些实施例中，也可以为预设个数的完整句子为预设的切分单位，可以为以 3 个句子为预设的切分单位。当然也可以 5 个句子，或者以 10 个句子为预设的切分单位，对知识文档进行切分，以得到由预设个数的完整句子构成的文档块。

应当理解，上述实施例仅为示例性说明，具体预设的切分规则可以根据用户需要灵活调整，并不以上述实施例给出的为限。只需保证基于预设的切分规则得到的各文档块均为可以正常表达语义的文档块即可。也就是说，文档块中所包括的文档内容不易过少，如果文档块中包括的文档内容过少，可能会存在切分后得到的文档块无法正常表达语义的情况。文档块中所包括的文档内容也不易过多，如果文档块中包括的文档内容过多，可能会存在一个文档多个不相关内容的情况，使得检索得到的信息中既包含有与问题相关的信息，也包含有与问题无关的信息，从而使得大语言模型生成的答案不准确。

在一些实施例中，为了保证后续各文档块调用的便捷性以及可靠性，可以在切分得到各文档块时，就为各文档块生成唯一标识，通过标识可以在多个文档块中唯一指示对应的文档块。各文档块的标识确定方式可以为：每个文档块从预设标识库中获取对应的标识，预设标识库中预设有大量的彼此之间互不相同的标识。预设标识库中已经被获取的标识将无法再被其他文档块获取。

各文档块的标识确定方式可以为：通过时间戳与随机数结合的方式来生成各文档块的唯一标识。例如使用切分获得各文档块的当前时间戳和随机数结合，来生成各文档块的唯一标识。

各文档块的标识确定方式可以为：根据雪花算法为各文档块生成唯一标识（雪花算法是一种分布式标识生成算法，只依赖于时间戳、机器标识和序列号来生成标识。其生成雪花算法的标识生成结果是一个 64 位的 long 型数字）。上述实施例仅为示例性说明，具体各文档块标识的生成方式可以根据用户需要灵活调整，并不以上述实施例给出的为限。

RAG 系统 120 是前述的检索增强生成架构下的系统。如前所述，RAG 系统 120 包

括检索和生成两个子任务。这两个子任务分别由不同的模型执行。RAG 系统 120 包括检索模型 121 和大语言模型 123。检索任务可以由检索模型 121 执行。生成任务可以由大语言模型 123 执行。RAG 系统 120 可以与源数据库 110 通信连接。检索模型 121 可以调用源数据库 110 中的知识文档，以根据用户输入的问题从源数据库 110 中检索出与问题相关的信息。大语言模型 123 可以基于问题以及与问题相关的信息生成问题的答案。

构建系统 130 是用于构建 RAG 系统 120 的测试集的系统。构建系统 130 可以存储有执行本说明书描述的 RAG 系统 120 的测试集的构建方法的数据或指令，并可以执行或者用于执行数据或指令。测试集可以用于对检索模型 121 的检索能力进行测试。如前所述，检索模型 121 可以根据用户输入的问题从源数据库 110 中检索出与问题相关的信息。检索能力可以用检索模型 121 所检索出的相关信息的准确度来衡量。因此，测试集中的数据可以作为检索模型 121 的标签数据。测试集中可以包括多个测试数据对。一个测试数据对可以包括一个测试问题以及测试问题对应的测试文档块。测试问题可以用于作为 RAG 系统 120 的输入问题。RAG 系统 120 中的检索模型 121 可以根据测试问题检索与测试问题相关的检索文档块。测试文档块可以作为测试问题对应的标准文档块。检索文档块与测试文档块的对比结果可以用来衡量检索模型 121 检索的准确度，进而评价检索模型 121 的检索能力。

构建系统 130 可以是单个的计算设备，也可以是由多个计算设备组成的集群系统。构建系统 130 中存储的执行本说明书描述的测试集的构建方法的数据或指令可以采用任何形式的系统架构。比如，分层架构、事件驱动架构、微核架构、微服务架构、或者云架构等等。

评估系统 150 是基于测试集对 RAG 系统 120 的检索能力进行评估的系统。评估系统 150 可以存储有执行本说明书描述的 RAG 系统 120 的评估方法的数据或指令，并可以执行或者用于执行数据或指令。评估系统 150 可以基于测试集中的测试数据对对检索模型 121 的检索能力进行评估。具体地，测试问题可以用于作为 RAG 系统 120 的输入问题。RAG 系统 120 中的检索模型 121 可以根据测试问题检索与测试问题相关的检索文档块。检索文档块与测试文档块的对比结果可以用来衡量检索模型 121 检索的准确度，进而评价检索模型 121 的检索能力。

评估系统 150 可以是单个的计算设备，也可以是由多个计算设备组成的集群系统。评估系统 150 中存储的执行本说明书描述的评估方法的数据或指令可以采用任何形式

的系统架构。比如，分层架构、事件驱动架构、微核架构、微服务架构、或者云架构等等。

需要说明的是，构建系统 130 和评估系统 150 可以是两个独立的系统，也可以集成在同一个系统中。

图 2 示出了根据本说明书的一些实施例提供的一种计算设备 200 的硬件结构示意图。构建系统的计算设备可以采用计算设备 200 的硬件结构。评估系统中的计算设备也可以采用计算设备 200 的硬件结构。如图 2 所示，计算设备 200 包括至少一个存储介质 230 和至少一个处理器 220。在一些实施例中，计算设备 200 还可以包括内部通信总线 210。在一些实施例中，计算设备 200 还可以包括通信端口 250。在一些实施例中，计算设备 200 还可以包括 I/O 组件 260。

内部通信总线 210 可以连接不同的系统组件，包括存储介质 230 和处理器 220。I/O 组件 260 支持计算设备 200 和其他组件之间的输入/输出。

存储介质 230 可以包括数据存储装置。数据存储装置可以是非暂时性存储介质，也可以是暂时性存储介质。比如，数据存储装置可以包括磁盘 232、只读存储介质（ROM）234 或随机存取存储介质（RAM）235 中的一种或多种。存储介质 230 还包括存储在数据存储装置中的至少一个指令集。指令集是计算机程序代码，计算机程序代码可以包括执行本说明书提供的检索增强生成系统的测试集的构建方法或者检索增强生成系统的评估方法的程序、例程、对象、组件、数据结构、过程、模块等等。

通信端口 250 用于计算设备 200 同外界的数据通讯。比如，计算设备 200 可以通过通信端口 250 连接网络 140。至少一个处理器 220 同至少一个存储介质 230 通过内部通信总线 210 通信连接。至少一个处理器 220 用以执行上述至少一个指令集。构建系统 130 运行时，至少一个处理器 220 读取至少一个指令集，并且根据至少一个指令集的指示执行本说明书提供的检索增强生成系统的测试集的构建方法。评估系统 150 运行时，至少一个处理器 220 读取至少一个指令集，并且根据至少一个指令集的指示执行本说明书提供的检索增强生成系统的评估方法。

处理器 220 可以执行检索增强生成系统的测试集的构建方法或者检索增强生成系统的评估方法包含的所有步骤。处理器 220 可以是一个或多个处理器的形式。处理器 220 可以发出执行指令。处理器 220 可以包括一个或多个硬件处理器，例如微控制器，微处理器，精简指令集计算机（RISC），专用集成电路（ASIC），特定于应用的指令集处理器

（ASIP），中央处理单元（CPU），图形处理单元（GPU），物理处理单元（PPU），微控制器单元，数字信号处理器（DSP），现场可编程门阵列（FPGA），高级 RISC 机器（ARM），可编程逻辑器件（PLD），能够执行一个或多个功能的任何电路或处理器等，或其任何组合。

仅仅为了说明问题，在本说明书中计算设备 200 中仅描述了一个处理器 220。然而，应当注意，本说明书中计算设备 200 还可以包括多个处理器，因此，本说明书中披露的操作和/或方法步骤可以如本说明书的由一个处理器执行，也可以由多个处理器联合执行。例如，如果在本说明书中计算设备 200 的处理器 220 执行步骤 A 和步骤 B，则应该理解，步骤 A 和步骤 B 也可以由两个不同处理器 220 联合或分开执行（例如，第一处理器执行步骤 A，第二处理器执行步骤 B，或者第一和第二处理器共同执行步骤 A 和 B）。

图 3 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建方法的流程图。如前，构建系统 130 中的计算设备 200 可以执行本说明书的检索增强生成系统的测试集的构建方法 P100。具体地，计算设备 200 中的处理器 220 可以读取存储在其本地存储介质中的指令集，然后根据指令集的规定，执行本说明书的检索增强生成系统的测试集的构建方法 P100。如图 3 所示，方法 P100 可以包括步骤 S110-S150。

S110：基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合。

一个测试文档组合内的文档块之间具有关联关系。所述关联关系可以是一个测试文档组合内的文档块之间存在相似性。所述相似性可以是具有相同或者相似的实体，或者内在存在潜在的联系（比如因果关系、连带关系、附属关系、递进关系、条件关系、承接关系、选择关系，等等）。不同测试文档组合内的文档块的数量可以相同也可以不同。一个测试文档组合内的文档块的数量可以是预设数量，也可以是不定量的数量。所述预设数量可以是一个预设值，也可以是一个预设的数值范围。

在步骤 S110 中，获取测试文档组合的方式可以有多种。在一些实施例中，构建系统可以通过知识抽取的方式获取文档块中的实体，并将具有相同实体的文档块作为一个测试文档组合。在一些实施例中，构建系统可以通过计算文档块两两之间的相似度将具有关联关系的文档块作为一个测试文档组合。在一些实施例中，构建系统还可以基于上述两种方式的组合获取测试文档组合。

在本说明书的实施例中，构建系统可以对多个文档块进行知识抽取，得到各文档块所包含的实体。实体可以是抽象的概念（如订单、事件、情感、人物、物体）。一个文档

块中可能包括一个实体，也可能包括多个实体。具体一个文档块包括实体的情况是根据真实的文档块的情况确定的，本说明书在此不做任何限制。随后，构建系统将具有相同实体的文档块作为一个测试文档组合。构建系统可以对通过知识抽取得到的每个实体执行文档块的关联操作。

由于具有相同实体的文档块很可能涉及相似的主题或领域，相同的实体在不同的文档块中的出现可能是因为它们都与该相似的主题紧密相关。其次，具有相同实体的文档块也可能在某些方面存在潜在的相关性。例如，两个文档块可能都涉及某个公司或某个产品，但各文档块例如是分别从不同的角度对相同实体进行描述的。在另一方面，实体是文档块语义内容的重要组成部分。因此，具有相同实体的文档块在语义上可能具有较高的相似性。所以通过将具有相同实体的文档块作为一个测试文档组合，可以保证测试文档组合与实体之间的强相关性。

具体地，对于一个实体，构建系统从多个文档块中确定出包含有当前实体的文档块，将这些文档块标记为具有关联关系的文档块，并将其作为一个测试文档块组合。在以文档块中的不同实体为组合目标时，对应的得到的测试文档组合可能不同。如前所述，一个文档块可以包括一个实体，也可以包括多个实体。当一个文档块包括多个实体时，同一个文档块可能只属于一个测试文档组合，也可能同时属于多个测试文档组合。举例说明，例如文档块 1 中包括多个实体，分别为实体 1、实体 2 和实体 3。若以文档块 1 中的实体 1 为组合目标，在确定文档块 3 和文档块 6 中均包括实体 1 之后，则针对实体 1 对应的测试文档组合可以为文档块 1-文档块 3-文档块 6。若以文档块 1 中的实体 2 为组合目标，在确定文档块 2 和文档块 8 中均包括实体 2 之后，则针对实体 2 对应的测试文档组合可以为文档块 1-文档块 2-文档块 8。若以文档块 1 中的实体 3 为组合目标，在确定文档块 5 和文档块 7 中均包括实体 3 之后，则针对实体 3 对应的测试文档组合可以为文档块 1-文档块 5-文档块 7。

也就是说，即使是对于同一文档块，在以该文档块中不同的实体作为组合目标时，可能对应的测试文档组合也不相同。应当理解，上述实施例虽然以每个测试文档组合中包括 3 个文档块为例进行描述。但是在确定一个测试文档组合时，具体一个测试文档组合中所包括的文档块的个数可以根据包含有相同实体的文档块的实际数量决定，并不以上述实施例给出的为限。一个测试文档组合中例如也可以包括 2 个文档块、或者 4 个文档块、5 个文档块等任意整数个数个文档块。

举例说明，在以实体为“NBA”为组合目标对各文档块进行检索时，例如可以检索

到均具有实体“NBA”的文档块 1、文档块 2 和文档块 3。其中文档块 1 可以为：“美国国家篮球协会（NBA）是北美洲的职业篮球联盟，由 30 支球队组成（美国 29 支，加拿大 1 支）。是美国和加拿大的主要职业体育联盟之一，被认为是世界上首屈一指的职业篮球通道”。文档块 2 可以为：“NBA 是美国篮球运动的活跃组织，被国际篮球协会称为美国篮球运动的国家管理机构，联盟的多个国际和个人球队办公室都设在曼哈顿城中城的总部。在北美，NBA 是继美国橄榄球联盟（NFL）和美国职业棒球大联盟（MLB）之后，收入第三富有的职业体育联盟，在全球排名前四”。文档块 3 可以为“波士顿凯尔特人队获得 NBA 总冠军最多，共 18 次，目前是联赛冠军，在 2024 年 NBA 总决赛中击败了达拉斯小牛队”。

在一些实施例中，构建系统 130 可以通过图谱的方式将具有关联关系的文档块进行关联。图 4 示出了根据本说明书的实施例提供的一种文档块关联图谱的示意图。图 4 中从左至右依次示出了 3 个测试文档组合。测试文档组合 1 中包含的实体为实体 1，多个包含实体 1 的文档块（文档块 1、文档块 3 和文档块 6）共同构成测试文档组合 1。测试文档组合 2 中包含的实体为实体 2，多个包含实体 2 的文档块（文档块 1、文档块 2、文档块 7 和文档块 8）构成测试文档组合 2。测试文档组合 3 中包含的实体为实体 3，多个包含实体 3 的文档块（文档块 1 和文档块 4）共同构成测试文档组合 3。可见，每个测试文档组合中包括的文档块的数量是根据包含有相同实体的文档块的实际数量决定，不同的测试文档组合中包括的文档块的数量可能不同。

或者，在另一些可能的实施例中，确定测试文档组合的方式可以为确定多个文档块两两之间的相似度。将两两之间的相似度高于预设阈值的文档块作为一个测试文档组合。应当理解，上述实施例仅为示例性说明，具体测试文档组合的确定方式是如何根据各文档块两两之间的相似度确定的，以及测试文档组合中所包括的文档块的个数，均可以根据用户需要灵活调整，并不以上述实施例给出的为限。

S130：基于大语言模型对测试文档组合进行问题抽取，确定测试文档组合对应的测试问题。

在一些可能的实施例中，对一个测试文档组合进行问题抽取（Question Extraction）也可以称为问题生成（Question Generation）。问题抽取的过程是指从给定的测试文档组合（如段落、文章、摘要等）中自动提取或生成与测试文档组合相关的问题。

在本说明书的实施例中，问题抽取的过程可以为大语言模型进行抽取的。大语言模

型可以是 RAG 系统中的大语言模型。在问题抽取的过程中，首先需要大语言模型对测试文档组合的内容进行深入的文本理解。理解的内容包括但不限于分析测试文档组合的主题、关键信息（实体信息）、实体关系（关系信息）等。理解的方式可以为使用自然语言处理（Natural Language Processing, NLP）技术，如词法分析、句法分析、语义分析等，来解析测试文档组合的结构和含义。对测试文档组合进行识别的内容包括但不限于：识别测试文档组合中的关键信息点（实体信息），如人名、地名、事件、概念等。这些关键信息点通常是测试问题生成的潜在候选者。

在一些实施例中，大语言模型可以直接基于针对测试文档组合的分析结果，生成测试文档组合对应的测试问题。或者，也可以将测试文档组合以及预先设定的提示工程模板一并输入至大语言模型中，大语言模型基于输入的测试文档组合和提示工程模板输出针对测试文档组合的测试问题。其中，提示工程模板用于引导大语言模型对测试文档组合进行问题总结。

其中，预先设定的提示工程模板可以为一系列预设的问题模板。也即通过预设的问题模板来引导大语言模型对测试文档组合进行问题总结，生成测试问题。这些模板例如可以根据测试文档组合中的不同信息类型进行调整。举例说明，对于测试文档组合中的实体为人名时，可以设计问题模板为“谁是 XXX？”。对于测试文档组合中的实体为事件，可以设计问题模板为“什么时候发生了 XXX？”或“为什么 XXX 会发生？”等。大语言模型可以使用识别出的关键信息点（目标实体）和预先设计好的问题模板，生成与测试文档组合内容相关的问题。

预先设定的提示工程模板可以是角色扮演提示模板。比如“假如你现在是一个【推理专家】，对以下文本进行推理，总结出一个与文本相关的问题。”

预先设定的提示工程模板可以是思维链提示模板。比如“请一步一步思考，对以下文本进行推理，总结出一个与文本相关的问题。”除此之外，在一些思维链提示模板中还可以包含一个或多个样例。

在生成问题之后，还可以对生成的问题进行筛选，去除与测试文档组合中的文本内容不相关、重复或质量低下的问题。

对生成的问题进行筛选的方式可以为使用机器学习或深度学习模型对生成的各问题进行打分或排序，以提高生成的问题的质量，以及与测试文档组合之间的相关性。最后可以对生成的问题进行人工评估或使用自动评估指标来衡量其质量和效果。根据评估

结果对问题抽取模型进行迭代和优化。应当理解，上述问题抽取过程仅为示例性说明，具体问题抽取过程的抽取方式以及抽取步骤均可以根据用户需要灵活调整，并不以上述实施例给出的为限。

S150：将测试问题与其对应的测试文档组合作为一组标签数据，构建测试集。

其中，测试集用于对检索增强生成系统的检索能力进行评估。为了提高对后续评估的准确性以及全面性，测试集包含有多组标签数据。一组标签数据包含一组测试数据对。一组测试数据对包括有一个综合性问题以及对应的由多个关联的文档块构成的测试文档组合。构建好的测试集能够对检索增强生成系统针对特定数据集的检索能力进行评估。相对于现有技术中的对检索增强生成系统的评估仅能针对一个单一问题对应一个文档块的评估方式。本说明书中的测试集的构建在对增强模型进行评估时，不仅可以评估检索增强生成系统根据问题检索单一文档的能力，还可以评估检索增强生成系统根据总结性的问题检索关联文档的能力。提高了对检索增强生成系统的检索能力评估的全面性。

在一些可能的实施例中，随着大语言模型的使用，源数据库中的知识文档在维护过程中可能会进行迭代和更新。这种情况下，为了保证测试集对于检索增强生成系统的检索能力评估的准确性，测试集也需要基于源数据库中的知识文档的迭代和更新进行迭代和更新。在本说明书的实施例中，构建系统还会响应于源数据库的更新事件，对测试集进行更新。

其中，源数据库的更新事件的触发方式可以为：在源数据库新增了预设个数个知识文档之后，则触发针对源数据库的更新事件。或者，可以为定时触发的源数据库的更新事件。其中，定时事件可以为预设的：每天、每周或者每半个月、或者每个月等任意天数，或每天、每周的任意时刻。或者，也可以为在源数据库新增了知识文档后，实时触发的针对源数据库的更新事件。应当理解，上述实施例仅为示例性说明，具体源数据库的更新事件的触发方式可以根据用户需要灵活调整，并不以上述实施例给出的为限。

采用本说明书提供的检索增强生成系统的测试集的构建方法，构建的测试集中的一组标签数据由一个测试问题和测试文档组合组成，一个测试文档组合包括相互关联的一个或多个文档块。而测试问题是基于测试文档组合生成的总结性问题。由此方式构建的测试集在对检索增强生成系统进行评估时，不仅可以评估检索增强生成系统根据问题检索单一文档的能力，还能评估检索增强生成系统根据总结性的问题检索关联文档的能力，从而提升了对检索增强生成系统的检索能力评估的全面性。并且在源数据库的更新事件

被触发时，还会对测试集进行更新，以根据更新后的测试集对检索增强生成系统进行进一步评估。

图 5 示出了根据本说明书的实施例提供的一种检索增强生成系统的测试集的构建的步骤示意图。如图 5 所示，测试集的构建步骤主要包括：对源数据库中的知识文档进行切分，以得到切分后的多个文档块。随后执行图 3 示出的 S110-S150。详细流程如下：

在获取到源数据库中存储的多个知识文档后，基于预设的切分规则，对多个知识文档进行切分，以得到切分后的多个文档块。其中，每个文档块均具有自己的唯一标识。标识的生成方式或确定方式可以为根据对各文档块的切分顺序，依次对各文档块进行标识的。或者，可以为根据对各文档块的切分时间，生成的各文档块对应的标识。或者，各文档块对应的标识也可以为根据其他方式确定的。对于各文档块对应表示的确定方式，本说明书在此不做任何限制，只需根据标识可以在多个文档块中唯一确定一个对应的文档块即可。

在本说明书的一个实施例中，结合图 5，以下述实施例为例对图 5 中的 S110 进行展开说明。如图 5 所示，例如对源数据库中的多个知识文档进行切分后，可以得到 N 个文档块。N 个文档块分别为文档块 1、文档块 2、文档块 3、……、文档块 N。基于切分得到的多个文档块，将具有关联关系的文档块作为测试文档组合。例如将具有关联关系的文档块 1-文档块 6-文档块 3 作为一个测试文档组合。当然，除了图 5 中的实例，也可以将具有关联关系的文档块 2-文档块 4-文档块 5 作为一个测试文档组合。具体组成方式根据实际情况确定，并不以上述实施例为限。

继续如图 5 所示，对图 5 中 S130 的展开说明如下：基于大语言模型分别对上述测试文档组合（文档块 1-6-3 和文档块 2-4-5）进行问题抽取，以生成上述测试文档组合对应的测试问题。文档块 1-文档块 6-文档块 3 进行问题抽取后得到对应的问题 1-6-3。文档块 2-文档块 4-文档块 5 进行问题抽取后得到对应的问题 2-4-5。将各测试问题与其对应的测试文档组合作为一组标签数据。多组标签数据共同构成测试集。例如：将问题 1-6-3 与测试文档组合 1-6-3 作为一组标签数据。将问题 2-4-5 与测试文档组合 2-4-5 作为一组标签数据。由多组标签数据构成的测试集用于后续对检索增强生成系统的检索能力进行评估。应当理解，上述实施例仅为示例性说明，具体测试集中包括的标签数据的内容，以及各测试文档组合中包括的文档块组合内容，均可以根据用户需要灵活调整，并不以上述实施例给出的为限。

图 6 示出了根据本说明书的实施例提供的一种检索增强生成系统的评估方法的流程图。如前，评估系统中的计算设备可以执行本说明书的检索增强生成系统的评估方法 P300。具体地，计算设备中的处理器可以读取存储在其本地存储介质中的指令集，然后根据指令集的规定，执行本说明书的检索增强生成系统的评估方法 P300。如图 6 所示，方法 P300 可以包括步骤 S310-S350。

S310：获取测试集。

测试集是基于前述的测试集的构建方法 P100 得到的，在此不再赘述。

S330：对于一组标签数据，将测试问题输入至检索增强生成系统的检索模型中，将检索模型输出的结果与测试问题对应的测试文档组合进行对比，得到对应的评估结果。

每一组标签数据对检索增强生成系统进行评估的方式可以为：将测试问题输入至 RAG 系统的检索模型后，得到检索模型输出的针对测试问题的回答文档。将回答文档与测试问题对应的测试文档组合进行对比。并根据上述对比结果对检索模型输出结果的准确性进行打分。评估结果可以是打分结果。评估结果的得分越高，则说明检索模型在这一组标签数据上的检索能力越高，检索准确度越高。如前所述，测试文档组合中可能包含多个文档块。而检索模型输出的回答文档中也可能包括多个文档块。回答文档和测试文档组合之间的对比可以是对比两个组合中相同的文档块的数量/比例以及不同的文档块的数量/比例。

在一些实施例中，回答文档和测试文档组合之间的对比可以是对比两个组合中相同的文档块的数量以及不同的文档块的数量。其中，为了提高对比效率，目标文档和测试文档组合中例如均只包括各文档块的标识。对比回答文档和测试文档组合的方式可以为对比两个组合中相同标识的数量。以根据对比结果中相同文档块的数量来确定对比结果的评分。例如根据相同文档块的数量占测试文档组合中所有文档块的数量的比例，以及预设满分分数来确定对比结果的评分。

举例说明：若当前回答文档与测试文档组合中相同文档块的数量为 X，测试文档组合中所有文档块的数量为 N，预设满分分数为 100 分，则针对当前检索模型输出的回答文档的准确性评分例如可以为 X/N*100，上述实施例仅为示例性说明，当然预设满分分数也可以为 5 分、10 分或任意分数，并不以此为限。

在另一些实施例中，可以根据当前回答文档与测试文档组合中相同文档块的数量直接对输出的回答文档的准确性进行打分。例如根据预设数量阈值与预设分值之间的对应

关系进行打分。举例说明：回答文档与测试文档组合中相同的文档块的数量为 1 时，对应的预设分值为 50 分。回答文档与测试文档组合中相同的文档数量为 2 时，对应的预设分值为 70 分。回答文档与测试文档组合中相同的文档块的数量为 3 时，对应的预设分值为 80 分。回答文档与测试文档组合中相同的文档块的数量例如大于或等于 4 时，对应的预设分值为 100 分（满分分数）。应当理解，上述相同文档块的数量与预设分值之间的对应关系仅为示例性说明，具体对应关系可以根据用户需要灵活调整，并不以上述实施例给出的为限。

当回答文档中出现与测试文档组合中相同的一个或多个文档块时，每个相同的文档块给予正分。而当回答文档中出现与测试文档组合中不同的一个或多个文档块时，每个不同的文档块给予负分。一组标签数据对应的评估结果可以是回答文档基于相同的文档块和不同的文档块所得的总分。

一个相同的文档块的分值的绝对值可以等于、小于或大于一个不同的文档块的分值的绝对值。比如，当一个相同的文档块的分值的绝对值可以等于一个不同的文档块的分值的绝对值时，当相同的文档块的数量大于不同的文档块的数量时，一组标签数据对应的评估结果为正分；当相同的文档块的数量小于不同的文档块的数量时，一组标签数据对应的评估结果为负分。

在一些实施例中，回答文档和测试文档组合之间的对比可以是对比两个组合中相同的文档块的比例以及不同的文档块的比例。相同的文档块的比例可以是回答文档中出现与测试文档组合中相同的文档块的数量在测试文档组合中的所有文档块数量的比例。不同的文档块的比例可以是回答文档中出现与测试文档组合中不同的文档块的数量在测试文档组合中的所有文档块数量的比例。相同的文档块给予正分。正分的大小与相同的文档块的比例成正比。相同的文档块的比例越高，正分越高。不同的文档块给予负分。负分绝对值的大小与不同的文档块的比例成正比。不同的文档块的比例越高，负分的绝对值越高。一组标签数据对应的评估结果可以是回答文档基于相同的文档块的比例和不同的文档块的比例所得的总分。

应当理解，上述实施例仅为示例性说明，具体确定检索增强生成系统中检索模型输出结果的准确性的评分的方式和规则可以根据用户需要灵活调整，并不以上述实施例给出的为限。

在另一些实施例中，评估结果也可以为通过/不通过。其中，评估结果是否指示通过，

是根据各组标签数据的评估评分确定的。例如在一组标签数据的评估评分高于预设阈值时，则确定上述一组标签数据的评估结果为通过。在一组标签数据的评估评分低于预设阈值时，则上述一组标签数据的评估结果确定评估结果为不通过。其中，预设阈值可以为用户预设的分数阈值。预设的分数阈值可以为用户根据需要，或根据历史经验预设的。可以为用户根据评分的满分分数，以及预设的比例阈值确定的。例如评分的满分分数为10分，预设的比例阈值为70%，则确定预设阈值为7分。在一组标签数据的评估分数低于7分时，则确定上述一组标签数据的评估结果为不通过。在一组标签数据的评估分数大于7分时，则确定上述一组标签数据的评估结果为通过。应当理解，上述实施例仅为示例性说明，具体评分的满分分数，以及预设的比例阈值均为可以灵活调整的，并不以上述实施例给出的为限。

在另一些实施例中，预设的分数阈值可能与场景有关。例如对于一些对检索模型的检索能力要求高的场景下，可以将预设的分数阈值设置的较高（或者将预设的比例阈值设置的较高）。对于一些对检索模型的检索能力要求低的场景下，可以将预设的分数阈值设置的较低（或者将预设的比例阈值设置的较低）。也就是说，在这种实施例中，评估结果是否指示通过，是根据各组标签数据的评估评分以及当前场景对应的预设的分数阈值（预设的比例阈值）综合确定的。例如评分的满分分数为100分，当前场景一对检索模型的检索能力要求较高，例如当前场景一对应的预设的分数阈值为90分。此时，在一组标签数据的评估评分为大于90分时，则确定上述一组标签数据的评估结果为通过。在一组标签数据的评估评分为小于90分时，则确定上述一组标签数据的评估结果为不通过。仍以评分的满分分数为100分为例进行说明，当前场景二对检索模型的检索能力要求较低，例如当前场景二对应的预设的分数阈值为60分。此时，在一组标签数据的评估评分为大于60分时，则确定上述一组标签数据的评估结果为通过。在一组标签数据的评估评分为小于60分时，则确定上述一组标签数据的评估结果为不通过。

可见在上述实施例中，对于同一组标签数据即使评估评分都相同，由于不通场景对检索能力的要求不同，对于同一组标签数据，在不同场景下对应的评估结果可能不相同。也即，评估结果与各组标签数据的评估评分以及场景对检索能力的要求综合确定。

在另一些实施例中，评估结果可以为评估等级，评估等级可以包括：高、中、低。预设的分数阈值可能包括有多个。确定各组标签数据的评估等级的方式可以为根据各组标签数据的评估评分与多个预设的分数阈值之间的关系，确定各组标签数据的评估等级。

以评分的满分分数为 10 分为例进行说明，例如第一分数阈值为 5 分，第二分数阈值为 7 分，第三分数阈值为 9 分。小于第一分数阈值的评估评分，则认为该评估评分的评估等级为低。大于第一分数阈值且小于第二分数阈值的评估评分，认为该评估评分的等级为中。大于第二分数阈值且小于第三分数阈值的评估评分，认为该评估评分的等级为高。在一组标签数据的评估评分为 5 分时，由于当前评估评分小于第一分数阈值，则确定上述一组标签数据的评估结果为低。在一组标签数据的评估评分为 8 分时，则确定上述一组标签数据的评估评分为中。在一组标签数据的评估评分为 10 分时，则确定上述一组标签数据的评估评分为高。应当理解，上述实施例仅为示例性说明，评估等级包括的等级数量，各等级的名称，具体各分数阈值也可以为根据满分分数和多个预设比例确定的，并不以上述实施例给出的为限。

S350：基于测试集对应的评估结果，确定目标评估结果。

在步骤 S330 中可以得到测试集中的多组标签数据对应的多个评估结果。评估系统可以基于多个评估结果得到检索模型最终的目标评估结果。在一些实施例中，目标评估结果可以为目标评估评分。如前所述，在评估结果为评估评分时，一个评估结果对应一个评估评分。目标评估结果可以是多个评估评分的总和。目标评估结果也可以是多个评估评分的加权和。目标评估结果也可以是多个评估评分的平均值。

或者，在评估结果为评估评分时，目标评估结果可能为通过/不通过。在确定目标评估结果之前，可以先根据多个评估结果的打分，确定一个总分数（基于测试集中的多组标签数据对应的评分分数确定的）。在总分数高于预设阈值时，则确定目标评估结果为通过。在总分数低于预设阈值时，则确定目标结果为不通过。

在一些实施例中，预设的分数阈值可能与场景有关。例如对于一些对检索模型的检索能力要求高的场景下，可以将预设的分数阈值设置的较高（或者将预设的比例阈值设置的较高）。对于一些对检索模型的检索能力要求低的场景下，可以将预设的分数阈值设置的较低（或者将预设的比例阈值设置的较低）。其中，预设的分数阈值是根据当前场景下的场景需求确定的，不同的场景需求可能对检索的准确度的要求不同，预设的分数阈值也可能不同。也就是说，针对同一检索模型以及同一测试集进行评估的过程中，即使输出的多个评估结果相同，在不同的场景需求下可能对应的检索模型是否通过评估的目标评估结果也不相同。

在另一些实施例中，目标评估结果可以为评估等级。评估等级可以包括：极高、高、

中、低、极低。预设的分数阈值可能包括有多个。确定目标评估结果的评估等级的方式可以为根据总分数与多个预设的分数阈值之间的关系，确定目标评估结果的评估等级。

在一些实施例中，总分数的确定方式可以为直接对测试集中的多组标签数据对应的评估评分进行求和。总分数也可以是对测试集中的多组标签数据对应的评估评分的加权和。总分数也可以是对测试集中的多组标签数据对应的评估评分的平均值。

在本说明书的实施例中，在针对一组标签数据进行评估阶段（步骤 S330 阶段）的预设阈值，可以与针对多组标签数据进行评估阶段（步骤 S350 阶段）的预设阈值相同，也可以不相同。具体不同阶段的预设阈值的设置方式可以根据用户需要灵活调整，并不以上述实施例给出的为限。

在测试集中的多组标签数据对应的评估结果为是否通过时，目标评估结果例如可以为根据评估结果指示通过的数量，测试集中的多组标签数据对应的评估结果的总数量以及预设阈值共同确定的。例如预设阈值可以设置为 80%，在指示通过的评估结果的数量与所有评估结果的总数量的比例大于 80%时，则确定目标评估结果为通过。在指示通过的评估结果的数量与所有评估结果的总数量的比例小于 80%时，则确定目标评估结果为不通过。

在测试集中的多组标签数据对应的评估结果为评估等级时，目标评估结果例如可以为根据评估结果指示的等级的数量，测试集中的多组标签数据对应的评估结果的总数量以及预设阈值确定的。例如预设阈值可以设置为 60%。在多个评估结果中，指示评估等级为高级的评估结果的数量与所有评估结果的总数量的比例大于 60%时，则确定目标评估结果对应的评估等级为高级。在多个评估结果中，指示评估等级为中级的评估结果的数量与所有评估结果的总数量的比例大于 60%时，则确定目标评估结果对应的评估等级为中级。在多个评估结果中，指示评估等级为低级的评估结果的数量与所有评估结果的总数量的比例大于 60%时，则确定目标评估结果对应的评估等级为低级。

在一些实施例中，为了保证目标评估结果的准确性。对于多个评估结果可以首先进行筛选，去除一些极端情况的评估结果。随后基于筛选后的多个评估结果，确定目标评估结果。这样的设置方式可以避免极端情况的评估结果的打分对目标评估结果造成较大影响。筛除方法例如可以为绝对阈值法：设定一个上限或下限，超过这些界限的值被认为是极端值。百分位法：将数据按大小排序，选取某个固定百分位值作为阈值，如第 95 个或第 5 个百分位数。

综上，本说明书提供的测试集的构建方法及系统、模型的评估方法及系统中，可以基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合。并基于大语言模型对测试文档组合进行问题抽取，以确定测试文档组合对应的问题。并将测试问题与对应的测试文档组合作为一组标签数据，构建测试集。这样的设置方式使得在使用增强检索生成模型之前，可以首先对源数据库中的多个文档块进行整理，以得到由多组标签数据构成的测试集。在一些可能的实施例中，测试集中可能包括多组问题和文档块之间一一对应的标签数据。测试集中也可能包括测试问题和测试文档组合一一对应的标签数据，并且测试文档组合是由多个文档块构成的。测试集的多样性可以保证对检索增强生成系统评估的全面性，使得本说明书提供的方法不仅可以评估检索增强生成系统根据问题检索单一文档的能力，还能评估检索增强生成系统根据总结性的问题检索关联文档的能力，从而提升对检索增强生成系统的检索能力评估的全面性，保证了后续检索增强生成系统回答问题的全面性以及可应用性。

本说明书另一方面提供一种计算机可读非暂时性存储介质，存储有至少一个指令集，用来进行检索增强生成系统的测试集的构建方法或检索增强生成系统的评估方法 任意方法的可执行指令。当至少一个指令集被处理器执行时，至少一个指令集指导处理器实施本说明书的检索增强生成系统的测试集的构建方法或检索增强生成系统的评估方法任意方法的步骤。在一些可能的实施方式中，本说明书的各个方面还可以实现为一种程序产品的形式，其包括程序代码。当程序产品在构建系统上运行时，程序代码用于使构建系统执行本说明书描述的方法 P100 的步骤。用于实现上述方法的程序产品可以采用便携式紧凑盘只读存储器（CD-ROM）包括程序代码，并可以在构建系统上运行。然而，本说明书的程序产品不限于此，在本说明书中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统使用或者与其结合使用。程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。计算机可读存储介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读存

储介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。可读存储介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF 等等，或者上述的任意合适的组合。可以以一种或多种程序设计语言的任意组合来编写用于执行本说明书操作的程序代码，程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。

综上，本说明书提供的检索增强生成系统的测试集的构建方法、检索增强生成系统的评估方法以及构建系统在对检索增强生成系统进行评估时，不仅可以评估检索增强生成系统根据问题检索单一文档的能力，还能评估检索增强生成系统根据总结性的问题检索关联文档的能力，从而提升对检索增强生成系统的检索能力评估的全面性。

上述对本说明书特定实施例进行了描述。其他实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者是可能有利的。

综上，在阅读本详细公开内容之后，本领域技术人员可以明白，前述详细公开内容可以仅以示例的方式呈现，并且可以不是限制性的。尽管这里没有明确说明，本领域技术人员可以理解本说明书需求囊括对实施例的各种合理改变，改进和修改。这些改变，改进和修改旨在由本说明书提出，并且在本说明书的示例性实施例的精神和范围内。

此外，本说明书中的某些术语已被用于描述本说明书的实施例。例如，“一个实施例”，“实施例”和/或“一些实施例”意味着结合该实施例描述的特定特征，结构或特性可以包括在本说明书的至少一个实施例中。因此，可以强调并且应当理解，在本说明书的各个部分中对“实施例”或“一个实施例”或“替代实施例”的两个或更多个引用不一定都指代相同的实施例。此外，特定特征，结构或特性可以在本说明书的一个或多个实施例中适当地组合。

应当理解，在本说明书的实施例的前述描述中，为了帮助理解一个特征，出于简化本说明书的目的，本说明书将各种特征组合在单个实施例、附图或其描述中。然而，这并不是说这些特征的组合是必须的，本领域技术人员在阅读本说明书的时候完全有可能

将其中一部分设备标注出来作为单独的实施例来理解。也就是说，本说明书中的实施例也可以理解为多个次级实施例的整合。而每个次级实施例的内容在少于单个前述公开实施例的所有特征的时候也是成立的。

本文引用的每个专利，专利申请，专利申请的出版物和其他材料，例如文章、书籍、说明书、出版物、文件、物品等，除了与本文件不一致或相冲突的、或者对权利要求的最宽范围具有限制性影响的文件，均可以通过引用结合于此，并用于现在或以后与本文件相关联的所有目的。此外，在任何材料对相关术语的描述、定义和/或使用与本文件对相关术语的描述、定义和/或使用之间存在任何不一致或冲突时，以本文件中的术语为准。

最后，应理解，本文公开的申请的实施方案是对本说明书的实施方案的原理的说明。其他修改后的实施例也在本说明书的范围内。因此，本说明书披露的实施例仅仅作为示例而非限制。本领域技术人员可以根据本说明书中的实施例采取替代配置来实现本说明书中的申请。因此，本说明书的实施例不限于申请中被精确地描述过的实施例。

权　利　要　求　书

1.一种检索增强生成系统的测试集的构建方法，包括：

基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合；

基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题；以及

将所述测试问题与其对应的测试文档组合作为一组标签数据，构建测试集，所述测试集用于对所述检索增强生成系统的检索能力进行评估。

2.如权利要求1所述的方法，其中，所述基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，包括：

对所述多个文档块进行知识抽取，得到文档块所包含的实体；以及

将具有相同实体的文档块作为一个测试文档组合。

3.如权利要求1所述的方法，其中，所述基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，包括：

确定所述多个文档块两两之间的相似度；以及

将两两之间的相似度高于预设阈值的文档块作为一个测试文档组合。

4.如权利要求1所述的方法，其中，所述基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，包括：

将所述测试文档组合以及预先设定的提示工程模板输入至所述大语言模型中，得到所述大语言模型输出的所述测试问题，所述提示工程模板用于引导所述大语言模型对所述测试文档组合进行问题总结。

5.如权利要求1所述的方法，其中，还包括：

响应于所述源数据库的更新事件，对所述测试集进行更新。

6.一种检索增强生成系统的评估方法，包括：

获取测试集，所述测试集是基于以下方式得到的：

基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，

基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，以及

将所述测试问题与其对应的测试文档组合作为一组标签数据，构建所述测试集；

对于一组标签数据，将所述测试问题输入至所述检索增强生成系统的检索模型中，

将所述检索模型输出的结果与所述测试问题对应的测试文档组合进行对比，得到对应的评估结果；以及

基于所述测试集对应的评估结果，确定所述检索模型的目标评估结果。

7.如权利要求 6 所述的方法，其中，所述基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，包括：

对所述多个文档块进行知识抽取，得到文档块所包含的实体；以及

将具有相同实体的文档块作为一个测试文档组合。

8.如权利要求 6 所述的方法，其中，所述基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，包括：

确定所述多个文档块两两之间的相似度；以及

将两两之间的相似度高于预设阈值的文档块作为一个测试文档组合。

9.如权利要求 6 所述的方法，其中，所述基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，包括：

将所述测试文档组合以及预先设定的提示工程模板输入至所述大语言模型中，得到所述大语言模型输出的所述测试问题，所述提示工程模板用于引导所述大语言模型对所述测试文档组合进行问题总结。

10.如权利要求 6 所述的方法，其中，所述检索增强生成系统的测试集的构建方法还包括：

响应于所述源数据库的更新事件，对所述测试集进行更新。

11.一种增强模型的训练系统，包括：

至少一个存储介质，存储有至少一个指令集，用于进行增强模型的训练方法；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并根据所述至少一个指令集的指示执行：

基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合；

基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题；以及

将所述测试问题与其对应的测试文档组合作为一组标签数据，构建测试集，所述测试集用于对所述检索增强生成系统的检索能力进行评估。

12.如权利要求 11 所述的系统，其中，为了将具有关联关系的文档块作为测试文档组合，所述至少一个处理器：

对所述多个文档块进行知识抽取，得到文档块所包含的实体；以及

将具有相同实体的文档块作为一个测试文档组合。

13. 如权利要求 11 所述的系统，其中，为了将具有关联关系的文档块作为测试文档组合，所述至少一个处理器：

确定所述多个文档块两两之间的相似度；以及

将两两之间的相似度高于预设阈值的文档块作为一个测试文档组合。

14.如权利要求 11 所述的系统，其中，为了基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，所述至少一个处理器：

将所述测试文档组合以及预先设定的提示工程模板输入至所述大语言模型中，得到所述大语言模型输出的所述测试问题，所述提示工程模板用于引导所述大语言模型对所述测试文档组合进行问题总结。

15.如权利要求 11 所述的系统，其中，所述至少一个处理器：

响应于所述源数据库的更新事件，对所述测试集进行更新。

16.一种检索增强生成系统的评估系统，包括：至少一个存储介质，存储有至少一个指令集，用于进行增强模型的训练方法；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，所述至少一个处理器运行时读取所述至少一个指令集，并根据所述至少一个指令集的指示执行：

获取测试集，所述测试集是基于以下方式得到的：

基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，

基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，以及

将所述测试问题与其对应的测试文档组合作为一组标签数据，构建所述测试集；

对于一组标签数据，将所述测试问题输入至所述检索增强生成系统的检索模型中，将所述检索模型输出的结果与所述测试问题对应的测试文档组合进行对比，得到对应的评估结果；以及

基于所述测试集对应的评估结果，确定所述检索模型的目标评估结果。

17.如权利要求 16 所述的系统，其中，为了基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，所述至少一个处理器：

对所述多个文档块进行知识抽取，得到文档块所包含的实体；以及

将具有相同实体的文档块作为一个测试文档组合。

18.如权利要求 16 所述的系统，其中，为了基于源数据库的多个文档块，将具有关联关系的文档块作为测试文档组合，所述至少一个处理器：

确定所述多个文档块两两之间的相似度；以及

将两两之间的相似度高于预设阈值的文档块作为一个测试文档组合。

19.如权利要求 16 所述的系统，其中，为了基于大语言模型对所述测试文档组合进行问题抽取，确定所述测试文档组合对应的测试问题，所述至少一个处理器：

将所述测试文档组合以及预先设定的提示工程模板输入至所述大语言模型中，得到所述大语言模型输出的所述测试问题，所述提示工程模板用于引导所述大语言模型对所述测试文档组合进行问题总结。

20.一种计算机可读非易失性存储介质，其中，所述计算机可读非易失性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如权利要求 1-10 中任一项方法。