**大型语言模型在信息检索中的应用：综述**

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen

**摘要**--作为信息获取的主要手段，信息检索（IR）系统，如搜索引擎，已深度融入我们的日常生活。这些系统还作为对话系统、问答系统和推荐系统的组成部分。信息检索技术经历了从基于术语的方法到结合先进神经模型的动态演进。尽管神经模型擅长捕捉复杂的上下文信号和语义细节，从而重塑信息检索的格局，但仍面临数据稀缺、可解释性差以及可能生成合理但不准确的结果等挑战。因此，信息检索的演进需要结合传统方法（如基于术语的稀疏检索方法）和现代神经架构（如具有强大语言理解能力的大型语言模型）。与此同时，ChatGPT和GPT-4等大型语言模型（LLMs）的兴起，以其卓越的语言理解、生成、泛化和推理能力，推动了自然语言处理领域的变革。近年来，研究人员尝试利用LLMs来优化信息检索系统。鉴于这一研究方向的快速发展，本文对现有方法进行整理，并提供深入见解。本综述探讨了LLMs在信息检索中的应用，包括查询重写、检索器、重排序器和阅读器等关键环节，并探讨了该领域未来的发展方向。

**关键词--**大型语言模型；信息检索；查询重写；重排序；阅读器；微调；提示工程

1. **引言**

信息获取是人类日常生活的基本需求之一。为了满足快速获取所需信息的需求，各种信息检索（Information Retrieval, IR）系统应运而生[1–4]。其中最具代表性的例子包括 Google、Bing 和百度等搜索引擎，它们作为互联网信息检索系统，能够根据用户的查询检索相关网页，并提供便捷高效的信息访问方式。值得注意的是，信息检索不仅限于网页检索。在对话系统（chatbots）[1, 5–8]中，例如微软小冰[2]、苹果Siri和谷歌助手，信息检索系统在从用户输入的语句中检索合适的响应方面发挥着关键作用，从而实现自然流畅的人机交互。同样，在问答系统[3, 9]中，信息检索系统用于选择关键信息，以有效回答用户的问题。此外，在图像搜索引擎[4]中，信息检索系统能够返回与用户输入查询相匹配的图像。随着信息量的指数级增长，学术界和工业界对开发高效的信息检索系统越来越感兴趣。

信息检索系统的核心功能是检索，其目标是确定用户查询与待检索内容（包括文本、图片、音乐等各种信息）之间的相关性。在本综述的范围内，我们主要关注文本检索系统，其中查询与文档的相关性通常通过匹配分数来衡量。由于信息检索系统需要在大规模数据存储库上运行，因此检索算法的效率至关重要。为了提升用户体验，检索性能可以从上游（查询重写）和下游（重排序与阅读）两个方面进行优化。作为上游技术，查询重写旨在优化用户查询，使其更有效地检索相关文档 [10,11]。近年来，随着对话式搜索的流行，这一技术受到了越来越多的关注。在下游阶段，重排序技术被用于进一步调整文档排名 [12–14]。与检索阶段不同，重排序仅在检索器已返回的有限相关文档集上进行，因此该阶段的重点在于提高检索性能，而非优化计算效率，从而可以采用更复杂的排序方法。此外，重排序还可以满足个性化 [15–18] 和多样化 [19–22] 检索等特定需求。在检索和重排序阶段之后，还引入了阅读组件，以对检索到的文档进行总结，并向用户提供简明的文档摘要 [23, 24]。传统的信息检索系统通常需要用户自行收集和整理相关信息，而在新一代信息检索系统（如New Bing）中，阅读组件已成为不可或缺的一部分，它能够优化用户的浏览体验，并节省宝贵的时间。

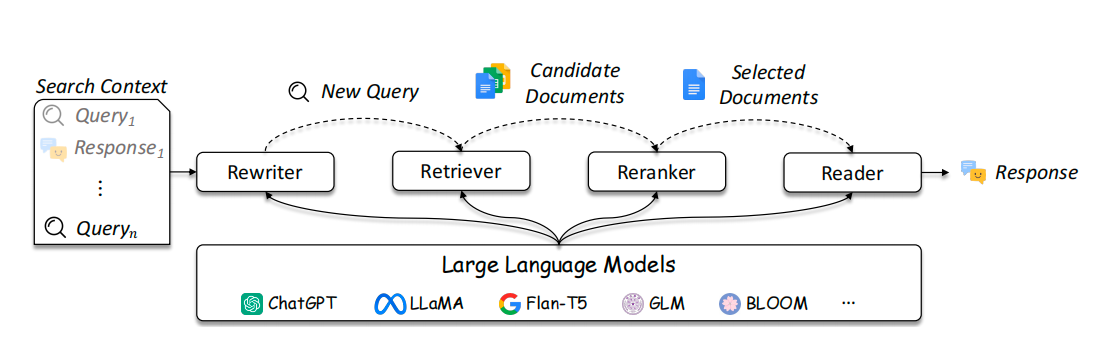


图 1. 将大型语言模型应用于信息检索的现有研究概述。大型语言模型可用于查询重写器、检索器、重排序器和阅读器

信息检索技术经历了动态的发展演进，从最初基于术语的方法逐步融合神经模型。最初，信息检索主要依赖基于术语的方法 [25] 和布尔逻辑（Boolean logic），侧重于基于关键词匹配进行文档检索。随着向量空间模型（Vector Space Models, VSM）[26] 的引入，这一范式逐渐发生转变，使得系统能够捕捉术语之间更细微的语义关系。这一进程进一步得到了统计语言模型（Statistical Language Models）[27, 28] 的推动，通过引入上下文和概率推理方法，提高了相关性估计的精确性。在此阶段，BM25 算法[29]发挥了重要作用，它通过考虑术语频率（Term Frequency, TF）和文档长度（Document Length）等因素，对相关性排名进行了优化，极大地提升了检索效果。信息检索领域的最新发展阶段标志着神经模型（Neural Models）[3, 30–32]的兴起。这些模型能够捕捉复杂的上下文线索和语义细节，从而重塑信息检索的格局。然而，神经模型仍然面临诸多挑战，如数据稀缺、模型可解释性差以及可能生成合理但不准确的检索结果。因此，信息检索的演进仍然需要在传统方法（如BM25算法的高效性）与现代神经架构（如神经网络的语义理解能力）之间找到平衡点，以充分发挥各自的优势。

近年来，大型语言模型（LLMs）已成为多个研究领域（如自然语言处理 [33–35]、推荐系统 [36–39] 甚至分子发现 [40]）中的变革性技术。这些LLMs主要基于Transformer结构，并在海量文本数据（包括网页、学术论文、书籍和代码）上进行大规模预训练。随着模型规模的不断扩大（包括参数数量和训练数据量），LLMs在多个方面表现出了卓越的能力。一方面，LLMs在语言理解和文本生成方面达到了前所未有的水平，使得其生成的文本更加接近人类语言表达，并且更加符合人类的意图。另一方面，大规模的LLMs在处理复杂任务时展现出显著的涌现能力，如泛化和推理能力[41]。值得注意的是，LLMs能够通过少量的任务示例或适当的指令，在无需额外微调的情况下，成功地解决新的任务。此外，诸如“上下文学习（In-Context Learning）”等先进技术的引入，显著提高了 LLMs的泛化能力，使其无需在特定下游任务上进行额外的微调即可获得卓越的任务表现 [34]。此外，得益于“链式思维（Chain-of-Thought）”等提示策略的加持，LLMs能够以逐步推理的方式生成输出，从而在复杂决策任务中展现更强的分析能力 [42]。利用LLMs强大的能力，无疑能够有效提升信息检索系统的性能。通过集成这些高级语言模型，IR系统可以为用户提供更精准的检索结果，从而重塑信息获取和检索的格局。

最初的努力是利用大型语言模型在开发新型红外系统中的潜力。值得注意的是，在实际应用方面，新必应旨在改善用户使用搜索引擎的体验，从不同的网页中提取信息，并将其浓缩为简洁的摘要，作为对用户生成查询的回应。在研究界，大型语言模型在IR系统的特定模块(如检索器)中被证明是有用的，从而提高了这些系统的整体性能。由于ll增强型红外系统的快速发展，有必要全面审查其最新进展和挑战。本综述深入探讨了LLMs与信息检索系统的交叉领域，涵盖了查询重写、检索器、重排序器和阅读器等关键方面（如图1所示）。这一分析增强了我们对LLMs在推动信息检索领域发展方面的潜力与局限性的理解。为支持本综述，我们创建了一个GitHub代码库，其中收集了有关LLM4IR的相关论文和资源。我们将持续更新该代码库，以收录最新的研究论文。此外，随着该领域的不断发展，我们也将定期更新本综述。我们注意到，已有多篇针对预训练语言模型（PLMs）、LLMs及其应用（如AIGC 或推荐系统）的综述[43–49]。其中，我们特别推荐文献 [49]，该综述系统而全面地涵盖了LLMs的多个重要方面。相比之下，我们的研究重点放在LLMs在信息检索系统中的开发与应用技术。此外，我们还发现了一篇探讨LLMs赋能信息检索系统机会的前瞻性论文[50]，该论文在未来研究方向上提供了有力的补充。

本综述的其余部分结构如下：第二节介绍了信息检索和LLMs的背景知识。第三、四、五、六节分别从查询重写、检索器、重排序器和阅读器四个核心组件的角度回顾了近年来的研究进展。随后，第七节探讨了未来研究的潜在方向。最后，在第八节，我们总结了本综述的主要研究发现。

1. **背景**
   1. **信息检索**

信息检索(Information retrieval, IR)是计算机科学的一个重要分支，其目的是从海量存储库中高效地检索与用户查询相关的信息。通常，用户通过以文本形式提交查询来与系统交互。随后，IR系统根据索引数据库对这些用户提供的查询进行匹配和排序，从而促进最相关结果的检索。

随着时间的推移，信息检索领域在各种模型的推动下取得了显著进展。其中，早期模型之一是布尔模型（Boolean Model），该模型使用布尔逻辑运算符来组合查询词，并检索满足特定条件的文档 [25]。基于“词袋”（bag-of-words）假设，向量空间模型（Vector Space Model）[26]将文档和查询表示为稀疏的基于词项的向量。该模型通过评估查询向量与文档向量之间的词汇相似性来进行相关性估计。此外，倒排索引的有效组织进一步提高了该模型的检索效率。更复杂的方法则引入了统计语言模型（Statistical Language Models），该模型用于估计词项出现的概率，并结合上下文信息，从而实现更加精确和具备上下文感知能力的检索 [27, 51]。近年来，神经信息检索（Neural IR）[30, 52, 53]范式在学术界受到了广泛关注。该范式利用神经网络的强大表示能力，捕捉查询与文档之间的语义关系，从而显著提升了检索性能。

研究人员发现，查询歧义性和检索效率等挑战会对IR系统的性能和有效性产生重要影响。针对这些挑战，研究人员将关注点聚焦于检索过程中至关重要的模块，以解决特定问题并实现相应的优化。这些模块在改善IR流程、提升系统性能方面发挥着不可忽视的作用。在本综述中，我们重点讨论了四个受大规模语言模型（LLMs）显著增强的关键模块。

**查询重写**（Query Rewriter）是信息检索（IR）系统中的一个关键模块，旨在提升用户查询的精确性和表达能力。该模块位于IR流程的早期阶段，承担着优化或修改初始查询的关键任务，以使其更准确地匹配用户的信息需求。作为查询重写的重要组成部分，查询扩展（Query Expansion）技术在提升查询表达能力方面起着核心作用，其中伪相关反馈（Pseudo Relevance Feedback）是最具代表性的方法之一。查询重写不仅在一般搜索场景中显著提高检索效果，还广泛应用于个性化搜索和对话式搜索等特定检索环境，进一步凸显了其重要性。

**检索器**（Retriever）通常用于信息检索（IR）系统的早期阶段，负责文档召回。检索技术的演进体现了人们对更高效、更精准方法的不懈追求，以应对日益增长的文本数据规模。在过去多年的IR系统实验中，经典的“词袋”模型BM25 [29] 以其稳健的性能和高效性证明了自身的价值。随着神经信息检索（Neural IR）范式的崛起，主流方法逐渐转向将查询和文档投射到高维向量空间，并通过内积计算相关性分数。这一范式转变使得IR系统能够更有效地理解查询与文档之间的关系，并利用向量表示捕捉语义相似性。

**重排序器**（Reranker）是检索流程中的另一个核心模块，其主要任务是对已召回的文档集进行精细化排序。与检索器侧重于效率与效果的平衡不同，重排序器更关注文档排名的质量。为了提高搜索结果的质量，研究人员探索比传统向量内积更复杂的匹配方法，以提供更丰富的匹配信号。此外，重排序器还支持个性化排序策略，例如个性化搜索和多样化搜索，以满足不同用户的检索需求。通过引入领域特定目标，重排序器能够提供定制化、精准的搜索结果，进一步优化用户体验。

**阅读器**（Reader）模块随着大规模语言模型（LLMs）的快速发展而成为不可或缺的组件。其核心能力在于理解实时用户意图，并基于已检索的文本生成动态响应，彻底革新了IR结果的呈现方式。与传统的返回候选文档列表相比，阅读器模块能够以更直观的方式组织答案，模拟人类自然的信息获取方式。

* 1. **大语言模型**

大语言模型（Language Models, LMs）旨在通过利用前文的上下文信息来计算词序列的生成概率，从而预测下一个单词的可能性。因此，借助特定的词选择策略（如贪心解码或随机采样），LMs能够高效生成自然语言文本。尽管LMs的主要目标在于文本生成，但最新研究[54]表明，多种自然语言处理（NLP）问题均可重新表述为文本到文本（text-to-text）的格式，从而通过文本生成来求解。这一发现使得LMs逐渐成为大多数文本相关问题的事实标准（de facto solution），广泛应用于各类自然语言处理任务。

语言模型（LMs）的演化可分为四个主要阶段，如先前的文献 [49] 所述。

最初，LMs 基于统计学习技术，被称为统计语言模型（Statistical Language Models）。这些模型通过马尔可夫假设（Markov Assumption），利用前面的单词预测下一个单词的概率，从而解决单词预测问题。随后，神经网络，特别是循环神经网络（RNNs），被引入用于计算文本序列的可能性，从而形成神经语言模型（Neural Language Models）。这些进展使得 LMs 不仅限于单纯的词序列建模，还能用于表示学习（Representation Learning）。在此之后，预训练语言模型（PLMs）开始兴起。例如，ELMo [55] 率先提出在大规模语料库上预训练一个双向 LSTM（biLSTM）网络，以学习上下文化的词表示，然后再针对特定的下游任务进行微调。类似地，BERT [56] 提出了基于 Transformer [57] 编码器的预训练方法，并设计了掩码语言模型（MLM）任务和下一句预测（NSP）任务，在大规模语料库上进行训练。这些研究开启了预训练-微调（Pre-training then Fine-Tuning）的新范式，并逐渐成为主流的方法。在这一方向上，许多生成式 PLMs（如 GPT-2 [33]、BART [58]和T5 [54]）被开发出来，专门用于文本生成任务，包括摘要、机器翻译和对话生成等。近年来，研究人员观察到扩大 PLMs 的规模（如模型大小或数据量）能够持续提高它们在下游任务中的表现，这一现象通常被称为规模法则（Scaling Law）[59, 60]。此外，大规模 PLMs 展现出许多较小规模模型不具备的能力，称为涌现能力（Emergent Abilities）[41]，使其能够应对更复杂的任务。因此，研究界将这些大规模 PLMs 称为大语言模型（Large Language Models, LLMs）。

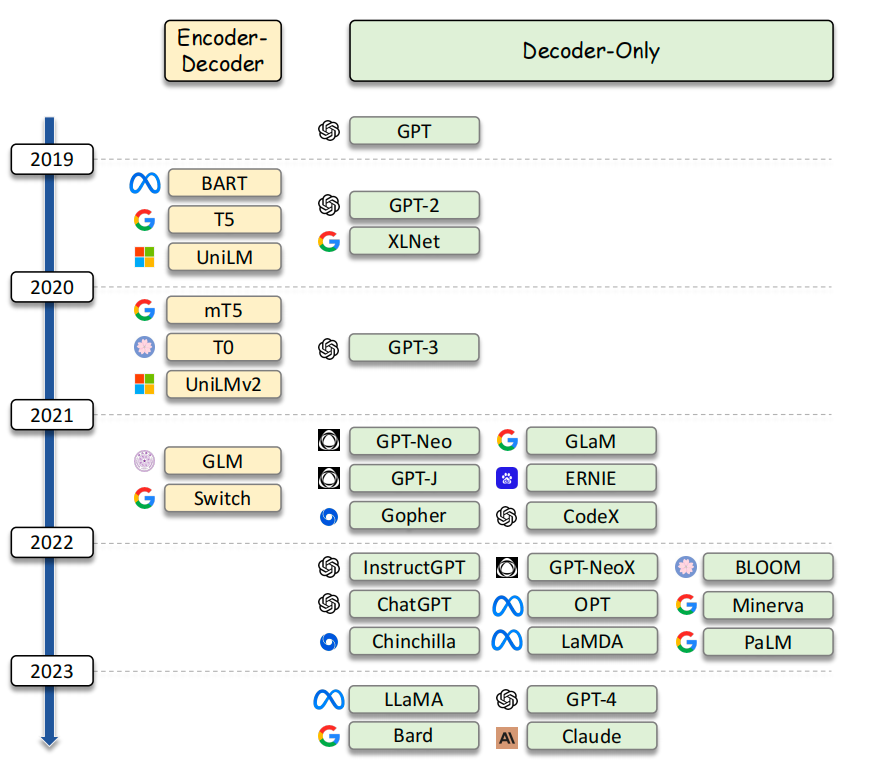


图 2. LLMs 的演变（编码器-解码器和仅解码器结构）

如图 2 所示，现有的大语言模型（LLMs）可根据其架构分为两大类：编码器-解码器（Encoder-Decoder）模型 [54, 58, 61–66] 和仅解码器（Decoder-Only）模型 [33–35, 67–77]。编码器-解码器（Encoder-Decoder）模型包含一个编码器（Encoder）组件，用于将输入文本转换为向量，并基于这些向量生成输出文本。例如，T5 [54] 就是一个典型的编码器-解码器模型，它将每个自然语言处理（NLP）问题转换为文本到文本（Text-to-Text）的格式，并以文本生成问题的方式进行求解。相比之下，仅解码器（Decoder-Only）模型（如 GPT）依赖于Transformer 解码器架构。该架构采用自注意力机制（Self-Attention），并结合对角注意力掩码（Diagonal Attention Mask），以从左到右的方式逐步生成单词序列。在GPT-3 [34] 取得成功后（该模型是首个参数规模超过 1000 亿的模型），多个受其启发的重要模型相继涌现，包括 GPT-J、BLOOM [75]、OPT [72]、Chinchilla [78] 和 LLaMA [35]。这些模型均遵循与 GPT-3 类似的 Transformer 解码器结构，并在不同组合的数据集上进行训练。

由于大型语言模型（LLM）拥有大量的参数，因此针对特定任务（例如信息检索）对其进行微调通常被认为是不切实际的。因此，人们建立了两种应用LLM的主要方法：上下文学习（ICL）和参数高效微调。ICL是LLM的一种新兴能力 [34]，它使LLM能够理解并根据提供的输入上下文提供答案，而不是仅仅依赖于其预训练知识。该方法仅需以自然语言形式描述任务描述和演示，然后将其作为输入提供给LLM。值得注意的是，ICL不需要参数调整。此外，通过采用思维链提示，可以进一步提高 ICL的有效性，该提示涉及多个演示（描述思维链示例）来指导模型的推理过程。ICL是将LLMs应用于IR的最常用方法。参数高效微调 [79–81] 旨在减少可训练参数的数量，同时保持令人满意的性能。例如，LoRA [79] 已被广泛应用于开源LLMs（例如 LLaMA和BLOOM）以实现此目的。最近，QLoRA [82] 被提出，通过利用冻结的 4 位量化LLM进行梯度计算来进一步减少内存使用。尽管参数高效微调在各种 NLP 任务中得到了探索，但其在 IR 任务中的实现仍然相对有限，代表着未来研究的潜在方向。

1. **查询重写器**

在传统的 IR 系统中，用户输入查询，系统返回与查询词匹配的文档列表。然而，原始查询通常很短或含糊不清，更容易出现词汇不匹配的问题。例如，如果用户在搜索引擎中输入查询“automobile”，他们期望找到关于汽车的信息。然而，搜索结果中大多数相关文档使用的是“vehicle”而不是“automobile”。由于搜索引擎主要依赖于精确的查询词，它可能无法检索到包含“vehicle”词语的最相关文档，从而导致词汇不匹配挑战。此外，在现代信息检索系统中，例如对话式搜索，查询重写任务扮演着更加重要的角色。

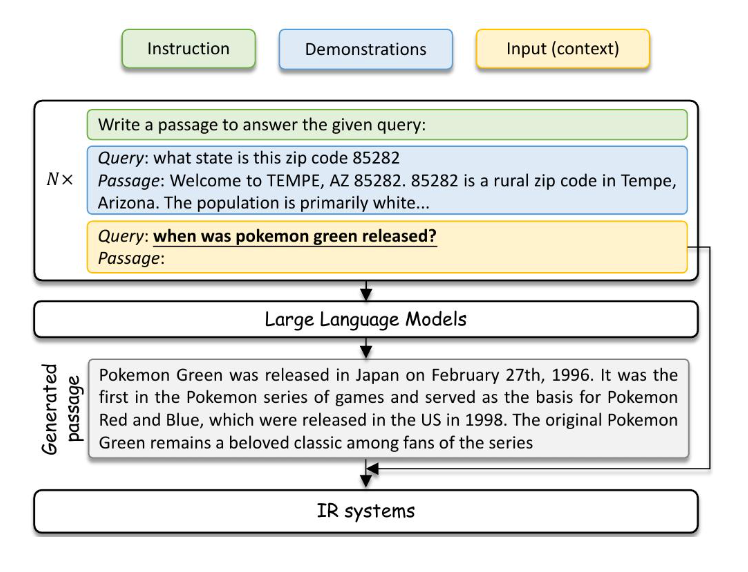


图 3. 基于 LLM 的即席搜索查询重写示例。该示例摘自 Query2Doc 论文 [87]。LLM 用于生成一段文字来补充原始查询，其中 N = 0 和 N > 0 分别对应于零样本和少样本场景。

传统的查询重写方法通过迭代地基于初始查询检索到的排名靠前的文档分析来细化用户的原始查询，从而解决这一挑战。这些方法通过迭代地细化查询表示并从文档集合中整合额外的信息，在提高检索准确性方面展现出了有效性。此类方法的显著例子包括 RM3 [83]、LCE [84]、KL 展开 [85] 和相关性建模 [86]。然而，这些方法通常依赖于预定义的规则或启发式方法，限制了它们完全捕捉用户意图细微差别的能力。

幸运的是，近年来大型语言模型 (LLM) 的发展为增强查询重写能力并克服传统方法的局限性提供了有希望的机会。这些强大的语言模型有可能利用海量的文本数据，并学习更细致入微的查询和文档表示。通过利用这些模型的能力，研究人员可以探索新的查询重写方法，以更好地满足用户复杂多样的信息需求。在接下来的部分中，我们将介绍在查询重写中使用大型语言模型的细节。

* 1. **重写目标**

查询重写通常服务于两种场景：即席检索，解决查询和文档之间的词汇不匹配问题；以及对话式搜索，根据不断发展的对话来细化和调整系统响应。在下一节中，我们将详细阐述这些追求背后的动机。

* + 1. **临时检索**

查询重写是即席检索中至关重要的一个步骤，因为它作为搜索漏斗的初始步骤，搜索过程的整体有效性高度依赖于查询重写结果的质量。在即席检索的背景下，查询重写的首要目标是检索与用户信息需求更匹配的文档集合。大型语言模型 (LLM) 在查询重写中的应用具有以下几个显著优势：

* 更好的语义理解。大型语言模型对语言语义有深刻的理解，使其能够更有效地捕捉查询的含义和上下文。相比之下，传统的查询重写方法主要依赖于词频的统计分析 [83–85]，这可能无法捕捉到查询的真实意图。
* 广博的知识。大型语言模型拥有广泛的知识，能够从各种概念、事实和信息中汲取知识。这种知识使它们能够利用对各种主题的理解，生成相关且符合语境的查询重写。
* 首次检索的独立性。传统的伪相关反馈 (PRF) 方法检索一组伪相关文档作为来源来细化原始查询。然而，伪相关反馈集中存在无关结果会引入噪声，并可能损害检索性能。相比之下，LLM 可以直接根据原始查询生成查询重写，这与首次检索无关，并防止潜在的噪声。

Query2Doc [87] 是一种基于大型语言模型 (LLM) 的查询重写器，它提示 LLM 根据查询生成相关段落（在某种程度上等同于生成答案）。随后，通过合并生成的段落来扩展原始查询。如图 3 所示，用户输入的原始查询为“口袋妖怪绿版什么时候发布”。LLM 通过使用少样本提示范式生成原始查询的相关段落。然后，将原始查询和生成的段落组合起来构建一个新的查询。检索模块使用这个新的查询来检索相关文档列表。值得注意的是，生成的段落包含额外的详细信息，例如“口袋妖怪绿版于2月27日在日本发行”，这在一定程度上有效地缓解了词汇不匹配问题。

* + 1. **对话式搜索**

另一方面，对话式搜索涉及用户和检索系统之间的动态交互，系统通过一系列来回交流来响应用户查询并澄清用户的的信息需求。在这种情况下，查询重写旨在根据不断发展的对话来完善和调整系统的响应。系统可能需要重写用户的查询或生成新的查询，以检索与当前对话相关的相关信息。

在大型语言模型 (LLM) 时代，在对话式搜索任务中利用 LLM 具有多种优势。首先，LLM 拥有强大的上下文理解能力，能够更好地理解用户在用户与系统之间多轮对话的语境下所表达的搜索意图。其次，LLM 表现出强大的生成能力，可以模拟用户与系统之间的对话，从而促进更强大的搜索意图建模。

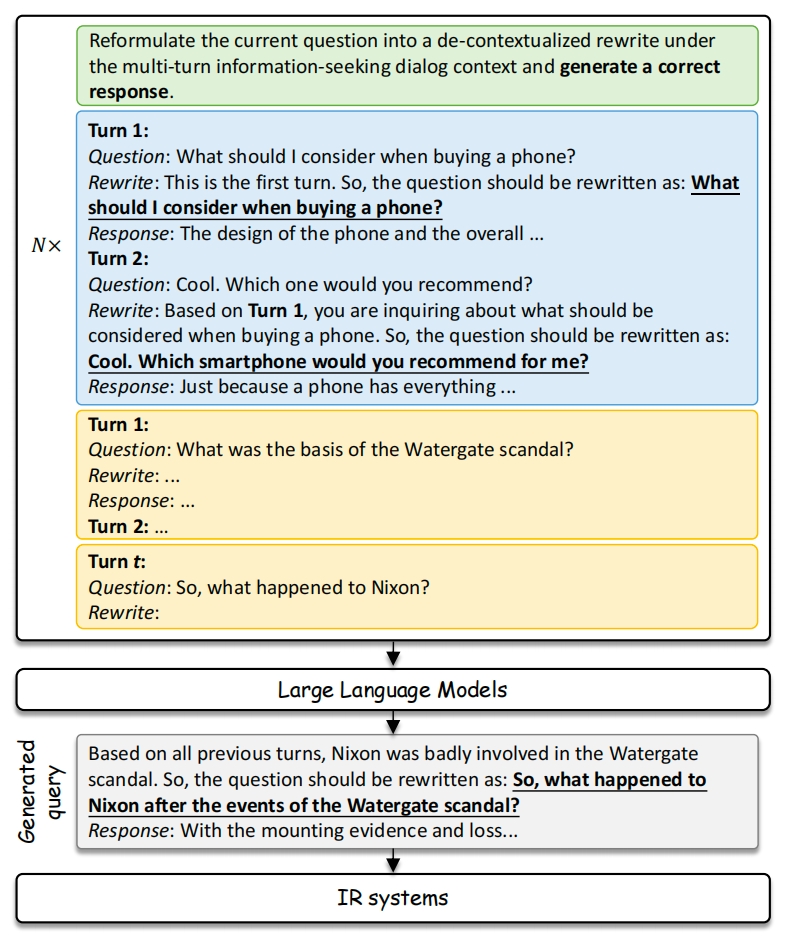


图 4. 基于LLM的对话式搜索查询重写示例。该示例来自LLMCS [88]。LLM 用于根据演示和之前的搜索上下文生成查询。需要生成额外的响应来改进查询理解。N = 0 和 N > 0 分别对应于零样本和少样本场景。

为了利用大型语言模型 (LLM) 在对话式搜索中的上下文理解和生成能力，提出了一种名为 LLMCS（“大型语言模型了解您的上下文搜索意图”）的框架 [88]。如图 4 所示，LLMCS提示LLM从多个角度生成查询重写和更长的假设系统响应。这些生成的输出随后被聚合到一个综合表示中，该表示能够稳健地捕捉用户的完整搜索意图。实验结果表明，将假设性响应与简洁的查询重写相结合，通过明确补充更合理的搜索意图，显著提高了搜索性能。

* 1. **数据资源**

查询重写方法通常需要补充语料库来丰富原始查询。LLM本身通过其参数包含世界知识。然而，这种通用知识可能不足以满足特定领域的需求，因此需要领域特定的语料库来提供更全面和领域特定的信息。在本节中，我们将分析两种方法：仅依赖模型中预先存在的知识的LLM方法和利用特定领域语料库来增强 LLM 能力的语料库增强型LLM方法。

* + 1. **大型语言模型的内在知识**

大型语言模型 (LLM) 能够在其参数中存储知识，这使其成为利用该知识进行查询重写的自然选择。作为基于LLM的查询重写领域的开创性工作，HyDE [89] 通过LLM根据给定的查询生成一个假设文档，然后使用密集检索器从语料库中检索与生成的文档相关的相关文档。Query2doc [87] 通过使用少量示例提示大型语言模型来生成伪文档，然后用生成的伪文档扩展查询。此外，还研究了不同提示方法和各种模型大小对查询重写的影響 [90]。为了更好地适应冻结检索器和基于大型语言模型的阅读器，使用了一个小型语言模型作为重写器，该模型使用强化学习技术进行训练，奖励由基于大型语言模型的阅读器提供 [91]。值得注意的是，所有这些工作都依赖于存储在大型语言模型 (LLM) 中的知识，而不是额外的语料库。尽管如此，实验结果表明，与传统的查询重写方法（例如RM3 [83]）相比，这些工作取得了显著的改进。

* + 1. **大型语言模型和文档语料库的内在知识**

尽管大型语言模型 (LLM) 表现出非凡的能力，但它们对特定领域的缺乏了解会导致生成幻觉或无关的查询。为了解决这个问题，最近的研究[92–95]提出了一种混合方法，通过文档语料库增强基于LLM的查询重写方法。例如，LLM 可能无法立即理解查询“ADHD”，但对检索到的文档的检查可以提供有价值的见解，表明它与一种称为“注意力缺陷/多动障碍”的疾病有关。

**为什么整合文档语料库？**整合文档语料库提供了几个显著的优势。首先，它通过微调特定主题领域的查询生成过程，提供领域特定知识，从而实现针对性和专业化的信息检索方法。其次，它通过整合从语料库中提取的事实信息，确保查询在可靠和可验证的知识基础上进行。第三，它通过用当代信息丰富查询来整合最新的概念，超越了大型语言模型中包含的知识，并确保了信息丰富度和及时性。最后，它通过利用相关文档作为补充资源来细化查询生成过程，有效减少无关信息的生成，并提高上下文相关输出的生成。

**如何整合文档语料库？**鉴于这些优势，人们提出了各种将文档语料库整合到基于LLM的查询重写中的范式，我们将其总结如下。

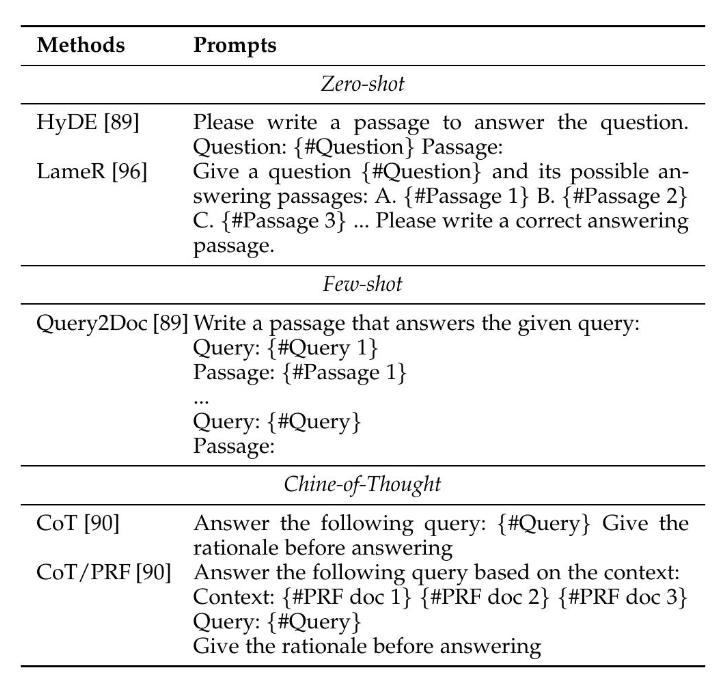
* 将查询和文档结合到LLM的输入中。在LLM时代，在提示中加入指令是一种实现特定功能的灵活方法。有几项研究利用了这种方法。GRM [92] 使用与相关性相关的样本估计来更准确地对重写项进行加权。对于查询意图分类任务，QUILL [93]证明了查询的检索增强为LLM提供了有价值的额外上下文，从而提高了理解能力。它提出了一种两阶段蒸馏管道，将LLM中的知识转移到小型模型中，以便更有效地重写查询。LameR [96] 提出了一种检索-重写-检索框架。它以BM25作为检索器，不使用任何带注释的查询-文档对，这表明LLM在零样本检索方面具有巨大潜力。此外，关于重写器是否应该出现在检索器之前没有严格的规则。InteR [94] 提出了一种支持搜索引擎和LLM之间多轮交互的框架，该框架允许搜索引擎使用LLM生成的知识集合扩展查询中的知识，并使LLM能够使用搜索引擎检索到的文档来增强提示的制定。
* 将来自大型语言模型 (LLM) 的生成式相关反馈 (GRF) 与来自文档语料库的伪相关反馈 (PRF) 相结合。GRF提供了首轮检索中不存在的外部上下文，而PRF将查询限制在目标语料库中包含的信息范围内。因此，结合生成式和伪相关反馈排序信号是实现两种反馈类别优势的有效方法。基于此动机，采用了一种加权融合方法 (GRF+PRF) 来结合GRF和PRF 的排序信号，以提高面向召回的有效性[95]。
  1. **生成方法**

利用大型语言模型 (LLM) 进行查询重写主要有三种范式。最常见的范式被称为提示方法，其中设计一个特定的提示或指令来引导语言模型生成期望的输出。提示方法为研究人员提供了重写过程的灵活性和可解释性。除了提示方法之外，在特定领域微调LLM和知识蒸馏也是有效的方法。LLM包含世界知识，这是一种通用的知识，但可能对特定领域无效。微调是指在特定数据集或任务上训练预训练的LLM，以增强其特定领域的有效性。检索增强会导致LLM推理的复杂性增加。知识蒸馏通过将LLM的知识转移到一些轻量级模型来缓解这个问题。在下一节中，我们将详细介绍这三种方法。

* + 1. **提示方法**

在大型语言模型 (LLM) 中，提示是指提供特定指令或上下文以指导模型生成文本的技术。提示的形式多种多样，从问题和不完整的句子到具体的指令。提示充当条件信号，影响模型的语言生成过程。因此，提示的有效性取决于提示的质量和设计。

表 1. 查询重写中不同提示方法的示例



目前流行的提示策略通常包含三类：零样本提示、少样本提示和思维链 (CoT) 提示 [42]。

* 零样本提示。零样本提示是指在没有接触过该领域或主题的训练样本的情况下，指示模型生成关于特定主题的文本。模型依靠其预先存在的知识和语言理解能力来为原始查询生成连贯且上下文相关的扩展词语。实验表明，零样本提示是一种简单但有效的查询重写方法[90, 94–98]。
* 少样本提示。少样本提示，也称为上下文学习，涉及为模型提供与所需任务或领域相关的有限示例或演示[87, 90, 97, 98]。这些示例充当一种显式指令，使模型能够将其语言生成适应当前的特定任务或领域。Query2Doc [87]提示模型根据排名数据集（例如MSMARCO [99]和NQ [100]）提供的某些查询-文档对，编写一段回答查询的文字。这项工作在一个提示上进行实验。为了进一步研究不同提示设计的的影响，[90]探索了八种不同的提示，例如提示大型语言模型生成查询扩展词而不是完整的伪文档，思维链提示等等。表1中有一些说明性的提示。这项工作比Query2Doc进行了更多实验，但实验表明，这种提示的效果不如Query2Doc中的提示。
* 链式思维提示。 CoT提示[42]是一种涉及迭代提示的策略，其中模型会收到一系列指令或部分输出[90, 92, 97]。在对话式搜索中，查询重写过程是多轮的，这意味着查询应该在搜索引擎和用户之间交互的过程中逐步细化。这个过程自然地与CoT过程一致。如4所示，工程师通过在每轮中添加一些指令来进行CoT过程，例如“基于所有之前的轮次，xxx”。而在即席搜索中，查询重写只有一轮。因此，CoT只能以简单粗略的方式完成。例如，如表 1 所示，研究人员在指令中添加“在回答之前给出理由”来提示LLM深思熟虑[90]。
  + 1. **微调方法**

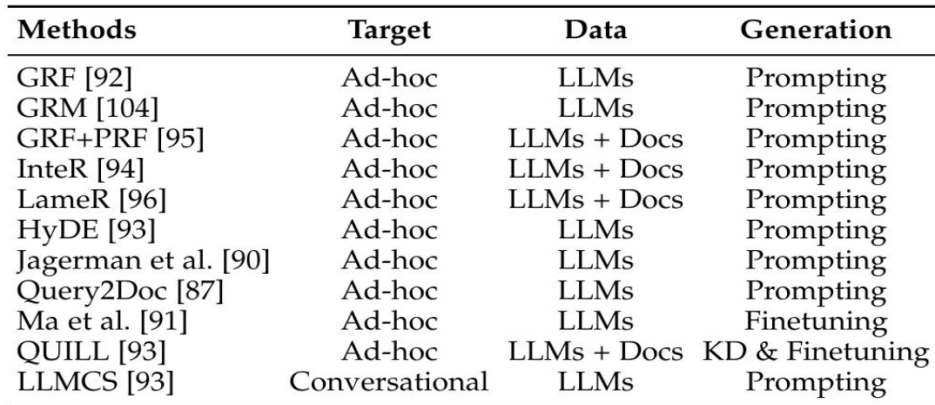
微调是另一种有效且普遍的范式，可以让LLM更好地适应特定领域。微调过程通常涉及使用预训练的LM，例如GPT-3，并在特定于目标领域的dataset上对其进行进一步训练。数据集可能包含查询、重写和相关标签，也可能通过结合人工专业知识和数据增强技术生成。在微调过程中，模型的参数会进行调整，以优化其在特定领域任务上的性能。

通过在特定领域数据上训练LLM，它可以学习特定领域的模式、术语和上下文，从而增强其生成高质量查询重写的能力。然而，微调LLM可能非常昂贵。有一些研究试图在查询重写任务上微调LLM。典型的研究微调大型语言模型 (LLM) 以生成给定查询[98]的相关文档，并将查询与生成的文档组合成新的查询。具体来说，它在NQ [100]和FEVER [102] 等下游数据集上，微调了具有7.7亿和30亿参数的FiD [101]（T5 的变体）。实验结果表明，使用550个token的输入长度微调770M FiD模型可以在单个Tesla V100 32GB GP上轻松完成，而训练3B T5模型则需要一个由八个Tesla V100或A100 GPU组成的更大集群。

* + 1. **知识蒸馏方法**

尽管基于 LLM 的方法在查询重写任务中取得了显著进步，但由于 LLM 的计算需求导致的巨大延迟，阻碍了它们在在线部署中的实际应用。为了缓解这一挑战，知识蒸馏已成为业界一种突出的技术。在 QUILL [93] 框架中，提出了一种两阶段蒸馏方法。这种方法需要使用检索增强型 LLM 作为教授模型，普通 LLM 作为教师模型，以及轻量级 BERT 模型作为学生模型。教授模型在两个使用了两个专门为查询意图理解而策划的大型数据集，即 Orcas-I [103] 和 EComm [93]。随后，采用两阶段蒸馏过程将知识从教授模型转移到教师模型，然后从教师模型转移到学生模型。实证结果表明，这种知识蒸馏方法优于简单地将模型大小从基础模型扩展到 XXL 模型，从而带来更大的改进。在一个“重写-检索-阅读”框架中 [91]，他们首先提示 LLM 重写查询，然后进行检索增强阅读。为了提高框架的有效性，采用了一个可训练的重写器，它被实现为一个小型语言模型，用于进一步修改搜索查询，使其符合冻结检索器和LLM阅读器的要求。为了微调重写器，采用了两步训练过程。首先使用伪数据进行监督预热训练。然后将检索-阅读管道建模为强化学习环境，并将重写器训练为策略模型，以最大化从管道性能中获得的奖励。

表2. 现有LLM增强型查询重写方法概述。 “Docs”和“KD”分别代表文档语料库和知识蒸馏。



* 1. **限制**

各种查询重写方法的实验一致地表明，增加生成的文档或答案数量对检索性能有积极影响[96]。虽然候选文档有助于呈现原始查询的补充上下文，但大多数现有工作通过重复或其他策略强调原始查询的突出地位[90, 94, 96]。据我们所知，目前缺乏专门用于评估查询重写的指标。现有的模型严重依赖于下游检索任务来评估查询重写器的有效性。如何直接判断重写后的查询是否反映了用户意图并有效地服务于特定任务仍然是一个未解决的问题。

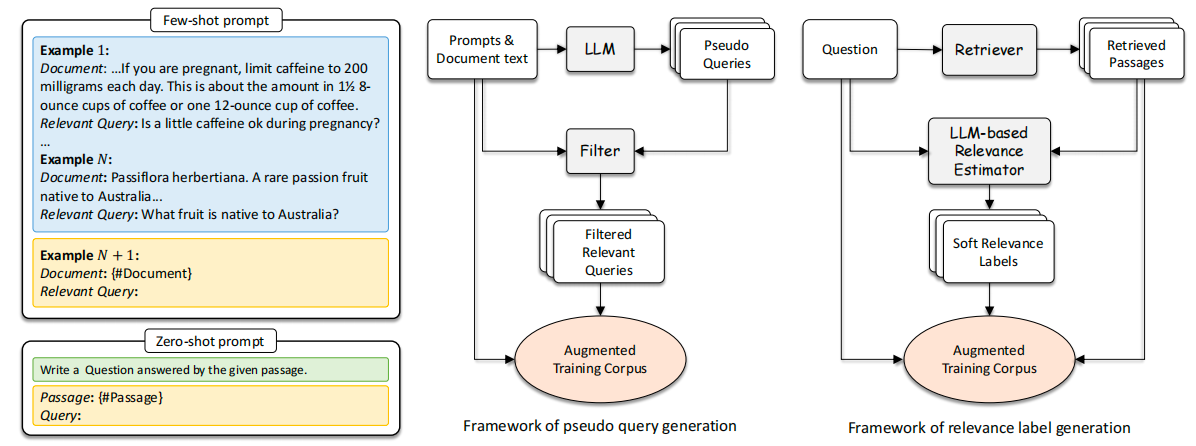


图 5. 基于LLM的数据增强在检索任务中的两个典型框架（右），以及它们的提示示例（左）。需要注意的是，相关性标签生成方法不将问题作为输入，而是将它们在检索到的段落上的生成概率视为软相关性标签。

1. **检索器**

在信息检索系统中，检索器充当首轮文档过滤器，用于收集与用户查询相关的广泛文档。鉴于信息检索系统中存在大量的文档，检索器在定位相关文档方面的效率对于维护搜索引擎性能至关重要。同时，检索器的高召回率对于最终结果的生成也很重要，因为检索到的文档会被馈送到排序器中，以生成最终结果供用户使用，这决定了搜索引擎的排序质量。

近年来，检索模型已从依赖统计算法[29]转向神经模型[3, 31]。后者在语义能力方面表现出色，擅长理解复杂的用户意图。神经检索器的成功依赖于两个关键因素：数据和模型。从数据角度来看，大量高质量的训练数据至关重要。这使检索器能够获取全面的知识和准确的匹配模式。此外，搜索数据的内在质量，即发出的查询和文档语料库，对检索性能有重大影响。从模型的角度来看，一个强表示性的神经网络结构允许检索器有效地存储和应用从训练数据中获得的知识。

不幸的是，一些长期挑战阻碍了检索模型的进步。首先，用户查询通常简短且含糊不清，这使得检索器难以准确理解用户的搜索意图。其次，文档通常包含冗长的内容和大量的噪音，这给编码长文档和提取相关信息以用于检索模型带来了挑战。此外，收集人工标注的相关性标签既费时又费钱。它限制了检索器的知识边界及其跨不同应用领域泛化的能力。此外，现有的模型架构，主要基于BERT [56]，存在固有的局限性，从而限制了检索器的性能潜力。近年来，大型语言模型 (LLMs) 在语言理解、文本生成和推理方面展现出非凡的能力。这促使研究人员利用这些能力来解决上述挑战，并帮助开发更优秀的检索模型。大体上，这些研究可以分为两类，即(1)利用LLMs生成搜索数据，以及(2)使用LLMs增强模型架构。

* 1. **利用大型语言模型生成搜索数据**

鉴于搜索数据的质量和数量，关于如何通过LLMs提高检索性能存在两种普遍的观点。第一种观点围绕搜索数据细化方法，重点在于重新表述输入查询以准确地呈现用户意图。第二个视角涉及训练数据增强方法，该方法利用LLM的生成能力来扩大密集检索模型的训练数据，特别是在零样本或少样本场景中。

* + 1. **搜索数据细化**

通常，输入查询由简短的句子或基于关键词的短语组成，这些短语可能存在歧义，并包含多种可能的用户意图。在这种情况下，准确地确定用户的特定意图至关重要。此外，文档通常包含冗余或噪声信息，这给检索器提取查询和文档之间的相关性信号带来了挑战。利用大型语言模型 (LLM) 强大的文本理解和生成能力为这些挑战提供了一个有希望的解决方案。迄今为止，该领域的研究工作主要集中在将LLM用作查询重写器，旨在优化输入查询以更精确地表达用户的搜索意图。第3节已经全面概述了这些研究，因此本节不再赘述。除了查询重写之外，一个引人入胜的探索方向是利用大型语言模型 (LLM) 通过细化冗长的文档来提高检索的有效性。这个引人入胜的领域仍有待进一步研究和发展。

* + 1. **训练数据增强**

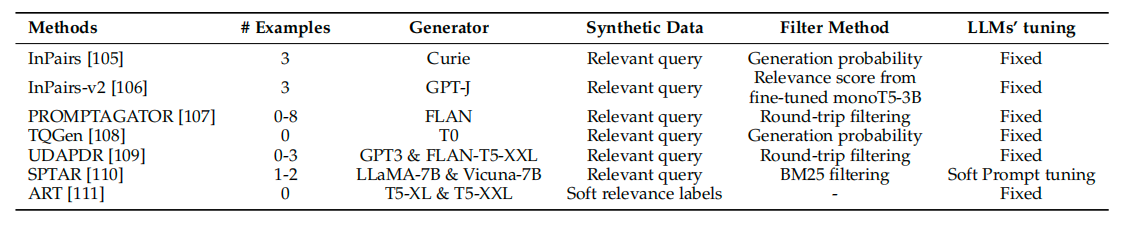
由于人工标注的经济和时间成本高昂，训练神经检索模型的一个常见问题是缺乏训练数据。幸运的是，大型语言模型 (LLM) 在文本生成方面的出色能力提供了一种潜在的解决方案。一个关键的研究重点在于设计策略，利用LLM 的能力生成伪相关信号，并为检索任务增强训练数据集。

**为什么我们需要数据增强？**先前对神经检索模型的研究集中在监督学习上，即使用来自特定领域的标记数据训练检索模型。例如，MS MARCO [99]提供了一个庞大的存储库，包含一百万段落、超过 200,000个文档和100,000 个带有人工标注的相关性标签的查询，这极大地促进了监督检索模型的发展。然而，这种范式本质上限制了检索器对来自其他领域的数据的泛化能力。检索模型的应用范围从自然语言问答到生物医学信息检索，并且为来自不同领域的数据标注相关性标签非常昂贵。因此，迫切需要零样本和少样本学习模型来解决这个问题[112]。在没有足够标签信号的情况下，提高模型在目标域中的有效性的一个常见做法是通过数据增强。

**如何将大型语言模型应用于数据增强？**在信息检索场景中，收集大量文档很容易。然而，收集真实用户查询并相应地标注相关文档是一项具有挑战性和成本高昂的任务。考虑到大型语言模型强大的文本生成能力，许多研究人员 [105, 106] 建议使用大型语言模型驱动的流程，根据现有集合创建伪查询或相关性标签。这些方法有助于构建相关的查询-文档对，从而扩大了检索模型的训练数据。根据生成数据的类型，有两种主流方法补充了基于LLM的检索模型数据增强，即伪查询生成和相关性标签生成。它们的框架在图5中可视化。接下来，我们将概述相关研究。

* 伪查询生成。鉴于文档数量众多，一个直接的想法是利用LLM来生成其对应的伪查询。inPairs [105]提供了一个这样的例子，它利用了GPT-3的上下文学习能力。该方法使用查询-文档对的集合作为演示。这些对与一个文档结合，作为输入提供给GPT-3，GPT-3随后会为给定的文档生成可能的相关查询。通过将相同的各种文档进行演示，可以轻松创建大量合成训练样本，并支持检索器在特定目标领域上的微调。为了提高这些合成样本的可靠性，采用了一个微调模型（例如，在MSMARCO [106] 上微调的monoT5-3B 模型）来过滤生成的查询。仅保留估计相关性得分最高的顶层对用于训练。这种“生成-过滤”范式可以以往返过滤的方式迭代进行，即首先在生成的样本上微调检索器，然后使用该检索器过滤生成的样本。重复这些类似EM的步骤直到收敛可以生成高质量的训练集[107]。此外，通过调整提供给LLMs的提示，它们可以生成不同类型的查询。这种能力允许更准确地模拟具有各种模式的真实查询[108]。在实践中，通过大型语言模型生成大量伪查询的成本很高。平衡生成成本和生成样本的质量已成为一个迫切的问题。为了解决这个问题，提出了 UDAPDR [109]，它首先使用大型语言模型为目标领域生成有限的合成查询集。这些高质量的示例随后被用作较小模型的提示，以生成大量查询，从而构建特定领域的训练集。值得注意的是，上述研究主要依赖于参数固定的固LLM。经验证明，优化大型语言模型(LLM)的参数可以显著提高其在下游任务中的表现。然而，这种追求受到计算资源需求过高的阻碍。为了克服这一障碍，SPTAR[110]引入了一种软提示微调技术，该技术仅在训练过程中优化提示的嵌入层。这种方法使LLMs能够更好地适应生成伪查询的任务，在训练成本和生成质量之间取得了良好的平衡。
* 相关性标签生成。在检索的一些下游任务中，例如问答，问题的集合也足够了。然而，连接这些问题与支持证据段落的相关性标签非常有限。在这种情况下，利用LLMs生成相关性标签的能力是一种很有前景的方法，可以增强检索器的训练语料库。最近的一种方法，ART [111]就是这种方法的例子。它首先检索与每个问题最相关的段落。然后，它使用LLM生成这些顶级段落条件下问题的生成概率。经过归一化处理后，这些概率作为检索器训练的软相关性标签。

表3. 基于大语言模型（LLMs）用于训练检索模型的现有数据增强方法比较。



此外，为了突出对应方法之间的异同，我们在表 3中展示了比较结果。该表从多个角度比较了上述方法，包括样本数量、所用生成器、生成合成数据的类型、用于过滤合成数据的技术以及是否对大型语言模型进行微调。本表旨在帮助更清晰地理解这些方法的概况。

* 1. **利用大语言模型增强模型架构**

利用大型语言模型 (LLM) 强大的文本编码和解码能力，可以比以前的小型模型 [56] 更精确地理解查询和文档。研究人员一直致力于利用LLM作为构建高级检索模型的基础。这些方法可以大体上分为两类，即基于编码器的检索器和生成式检索器。

* + 1. **基于编码器的检索器**

除了数据的数量和质量外，模型的代表性能力也极大地影响了检索器的效力。受大型语言模型 (LLM) 在自然语言建模和理解方面的出色能力的启发，一些研究者[113–115]利用LLM作为检索编码器，并研究了模型规模对检索器性能的影响。

**通用检索器。**由于检索器的有效性主要依赖于文本嵌入的能力，文本嵌入模型的演进往往对检索器的发展进程产生重大影响。在大型语言模型时代，OpenAI做出了开创性的工作[113]。他们将相邻的文本片段视为正样本对，以促进一系列文本嵌入模型（表示为cpt-text）的无监督预训练，这些模型的大小从300M到175B个参数不等。在MS MARCO [99]和BEIR [112]数据集上的实验表明，更大的模型规模有可能在文本搜索任务的无监督学习和迁移学习中取得更好的性能。然而，从头开始预训练大型语言模型对于大多数研究人员来说成本过高。为了克服这一限制，研究人员使用Base、Large、XL和XXL 等更小尺寸的T5家族模型来初始化双编码器的模型参数，然后在检索数据集上进行微调[114]。实验结果再次证实，更大的模型尺寸可以带来更好的性能，尤其是在零样本设置中。

**任务感知检索器。**虽然上述研究主要集中于使用LLM作为文本嵌入模型用于下游检索任务，但当集成特定任务指令时，检索性能可以得到极大提升。例如，TART [115]设计了一种任务感知检索模型，在问题之前引入特定任务指令。本指令包含对任务意图、领域和所需检索单元的描述。例如，假设任务是问答，一个有效的提示可能是“检索一篇回答此问题的维基百科文本。{question}”。这里，“维基百科”（领域）表示预期检索文档的来源，“文本”（单元）表示要检索的内容类型，而“回答此问题”（意图）则说明了检索文本与问题之间的预期关系。这种方法可以利用大型语言模型强大的语言建模能力和广泛的知识，精确地捕捉用户在各种检索任务中的搜索意图。

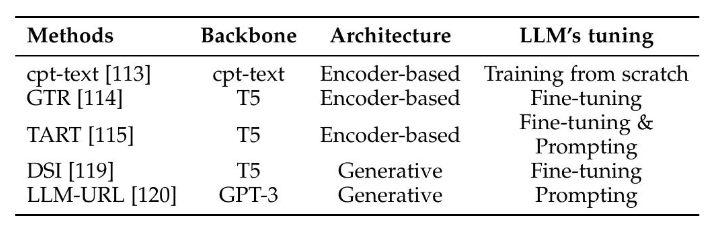
* + 1. **生成式检索器**

传统的 IR 系统通常遵循“索引-检索-排序”范式，根据用户查询定位相关文档，这在实践中已被证明是有效的。然而，这些系统通常由三个独立的模块组成：索引模块、检索模块和重排序模块。因此，对这些模块进行整体优化可能具有挑战性，并可能导致次优的检索结果。此外，这种范式需要额外的空间来存储预先构建的索引，从而进一步加重存储资源的负担。最近，基于模型的生成式检索方法[116–118]应运而生，旨在解决这些挑战。这些方法摆脱了传统的“索引-检索-排序”范式，而是使用统一模型直接生成与查询相关的文档标识符（即DocID）。在这些基于模型的生成式检索方法中，文档语料库的知识存储在模型参数中，无需为索引分配额外的存储空间。现有的方法已经探索了通过微调和提示大型语言模型来生成文档标识符[119, 120]。

**微调大型语言模型。**鉴于大型语言模型包含大量的世界知识，利用它们构建基于模型的生成式检索器是直观的。DSI [119]是一种典型的将预训练的T5 模型微调到检索数据集上的方法。该方法涉及直接对查询进行编码并解码文档标识符以执行检索。他们探索了多种生成文档标识符的技术，并发现构建语义结构化标识符可以产生最佳结果。在这种策略中，DSI 应用层次聚类根据文档的语义嵌入对文档进行分组，并根据其层次分组为每个文档分配一个语义 DocID。为了确保输出的DocID有效且代表语料库中的实际文档，DSI 使用所有DocID构建了一个字典树，并在解码过程中利用约束束搜索。此外，这种方法观察到，表明更大的语言模型会导致性能提升的扩展定律也适用于生成式检索器。

**提示大型语言模型。**除了微调大型语言模型以进行检索之外，人们还发现大型语言模型（例如，GPT系列模型）可以通过几个上下文演示直接为用户查询生成相关的网页URL [120]。大型语言模型的这种独特能力被认为源于它们在训练中接触了各种HTML资源。因此，LLMs可以自然地充当生成式检索器，直接生成文档标识符以检索与输入查询相关的文档。为了实现这一点，提出了一种LLM-URL[120]模型。它利用GPT-3 text-davinci-003模型来生成候选URL。此外，它设计正则表达式从这些候选者中提取有效的URL，以定位检索到的文档。

表 4. 基于大型语言模型的检索器比较。



为了全面理解该主题，表4总结了上述基于LLM的检索器的共同特征和独特特征。

**4.3限制**

尽管在LLM增强的检索方面已经做出了一些努力，但仍有许多领域需要更详细的调查。例如，检索器的一个关键要求是快速响应，而现有LLM的主要问题是模型参数巨大，推理时间过长。解决LLMs的这一局限性以确保检索器的响应时间是一个关键任务。此外，即使使用LLMs来增强数据集（推理时间要求较低的上下文），LLM生成的文本与真实用户查询之间潜在的错配也会影响检索效果。此外，由于大型语言模型通常缺乏领域特定知识，因此需要在特定任务数据集上进行微调，才能应用于下游任务。因此，开发高效的策略来微调这些具有大量参数的大型语言模型成为一个关键问题。

1. **重排序器**

重排序器，作为信息检索中的第二阶段文档过滤器，旨在根据文档与用户查询的相关性对检索器（例如，BM25）检索到的文档集进行重新排序。在大型语言模型的应用方面，现有的重排序方法可以分为三个主要范式：微调大型语言模型进行重排序，提示大型语言模型进行重排序，以及利用大型语言模型进行数据增强训练。这些范式将在接下来的部分中详细阐述。回顾一下，我们将使用术语“文档”来指代在一般IR场景中检索到的文本，包括段落（例如，MS MARCO段落排序数据集[99]中的段落）。

* 1. **微调大型语言模型以重排序**

微调是将预训练的LLM应用于重排序任务的重要步骤。由于在预训练过程中缺乏对排序的意识，LLM无法适当地衡量查询-文档的相关性，也无法完全理解重排序任务。通过在特定任务的排序数据集上微调LLM，例如MS MARCO段落排序数据集[99]，该数据集包含相关性和不相关性的信号，LLM可以调整其参数以在重排序任务中获得更好的性能。一般来说，现有的LLM微调策略可以分为两种主要方法：（1）将LLM微调为生成模型，以及（2）将LLM微调为排序模型。

* + 1. **将大型语言模型微调为生成模型**

在该领域，现有研究通常将文档排序表述为生成任务，并通过生成损失来优化大型语言模型 [13, 121, 122]。具体来说，重新排序模型通常被微调以生成单个标记，例如“真”或“假”，给定查询和文档。在推理过程中，查询-文档相关性得分是根据生成token的logit值来确定的。例如，可以微调T5模型以生成相关或不相关的查询-文档对的分类token[13]。在推理时，对“true”和“false”token的logit应用softmax函数，相关性得分计算为分配给“true”token的概率。以下方法[121]采用基于T5模型的多视角学习方法。该方法同时考虑两个任务：为给定的查询-文档对生成分类标记，以及根据提供的文档生成相应的查询。DuoT5 [122]将三元组 作为 T5 模型的输入，并进行微调，以生成标记“true”，如果文档 比文档 更相关于查询 *qi*，否则生成“false”。在推理过程中，对于每个文档 *di*，它枚举所有其他文档 *dj*，并使用全局聚合函数为文档 生成相关性分数 *si*（例如，*si*= *Pj pi,j*，其中 *pi,j*表示以作为模型输入时生成“真”的概率）。

* + 1. **微调大型语言模型作为排序模型**

尽管将大型语言模型微调为生成模型的方法优于一些强大的排序基线，但它们对于重排序任务并非最佳选择。 这主要源于两个原因。 首先，人们普遍期望重排序模型为每个查询-文档对生成一个数值相关性分数，而不是文本标记。其次，与生成损失相比，使用排序损失（例如，RankNet [123]）优化重排序模型更合理。尽管一些预训练模型，如BERT，已被用于文档重排序，但使用基于seq2seq的大型语言模型（LLM），如 T5-3B，进行重排序任务尚未得到充分研究。最近，RankT5 [124] 直接计算查询-文档对的相关性分数，并使用“成对”或“列表”排序损失优化排序性能。通过用更大规模的 T5 模型替换基础尺寸的 T5 模型，可以实现潜在的性能提升。

* 1. **利用大型语言模型进行重排序**

随着大型语言模型（LLM）规模的扩大（例如，参数超过 100 亿），微调重排序模型变得越来越困难。为了解决这一挑战，最近的研究尝试利用LLM 强大的指令能力，通过提示策略直接增强文档重排序。一般而言，这些用于重排序的提示策略可以分为三类：点式、列表式和成对方法。后续部分将对这些策略进行全面探讨。

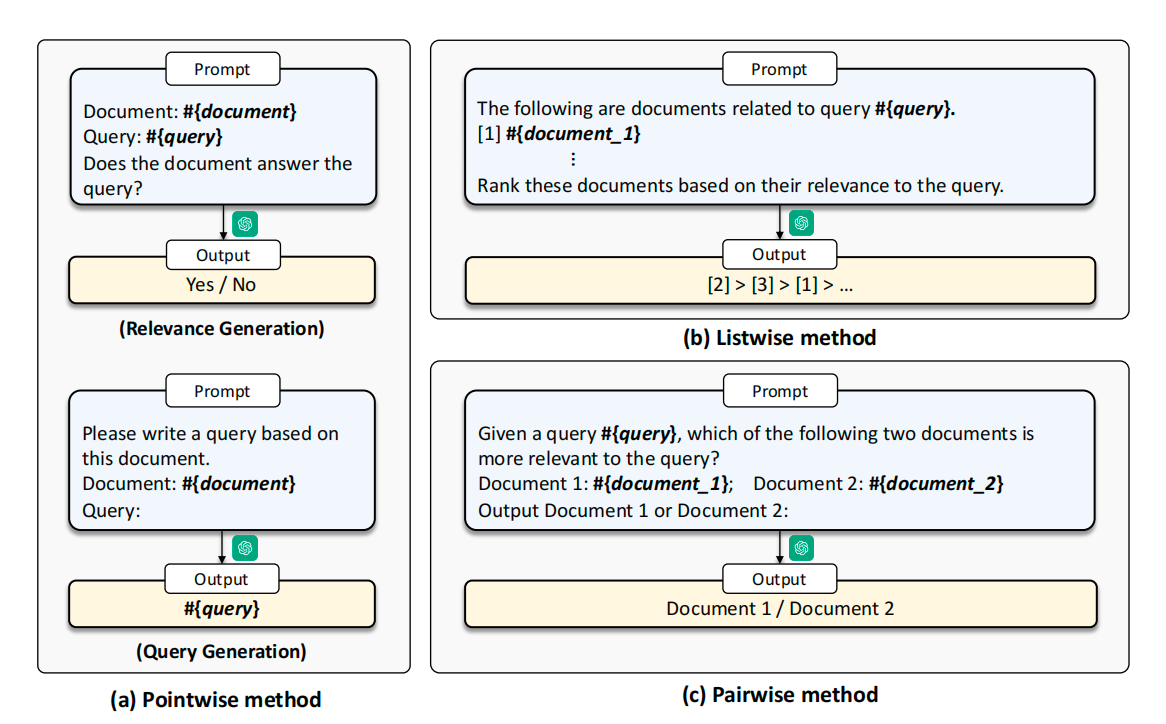
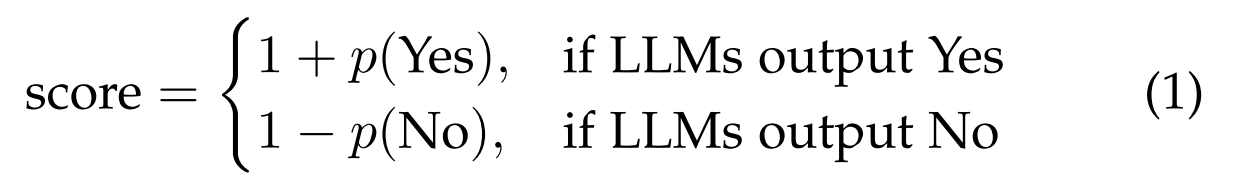


图 6. 三种基于提示的重排序方法：(a) 点式方法，包括相关性生成（上）和查询生成（下），(b) 列表式方法，以及 (c) 成对方法。

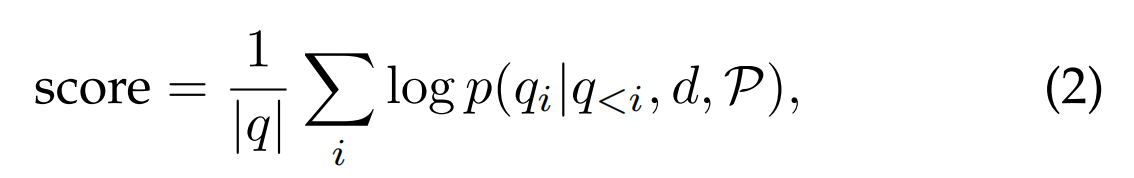
* + 1. **点态方法**

逐点方法测量单个文档与查询之间的相关性，可分为两种类型：相关性生成 [125] 和查询生成 [126, 127]。图 6 (a) 上部显示了基于给定提示的相关性生成，其中 LLMs 在文档与查询相关时输出“是”，否则输出“否”。大型语言模型生成“是”或“否”的概率用于计算查询-文档相关性分数 s，如下所示：



其中 p(Yes/No) 代表“Yes”或“No”标记的生成概率。

关于查询生成，图 6 (a) 下部展示了一个示例，其中查询-文档相关性得分由基于文档生成实际查询词元的平均对数似然决定：



其中 |q| 表示查询 q 的词元数量，d 表示文档，P 表示提供的提示。然后根据文档的相关性得分对文档进行排序。现有的基于查询生成的 方法 [126] 主要依赖于手工制作的提示（例如，“请根据此文档编写一个查询”），这可能不是最佳选择。由于提示是指导大型语言模型 (LLM) 执行各种自然语言处理 (NLP) 任务的关键因素，因此识别最佳提示以获得最佳模型性能至关重要。沿着这条线，提出了一种离散提示优化方法 Co-Prompt [127]，用于在重排序任务中生成更好的提示。

* + 1. **列表法**

逐点方法使用输出标记的对数概率来计算查询-文档的相关性得分，而这通常在 LLM API（例如 ChatGPT 和 GPT4）中不可用。最近，提出了一种直接对文档列表进行排序的列表式方法 [128, 129] 来解决这个问题。这些方法将查询和文档列表插入提示中，并指示LLMs输出重新排序的文档标识符（见图6 (b)）。考虑到LLMs的有限输入长度，这些方法还采用滑动窗口策略，每次重新排序候选文档的子集[128]。基于多个大型语言模型的广泛实验表明，基于 GPT-4 的方法取得了具有竞争力的结果，甚至在某些信息检索基准测试中超越了一些监督方法。

* + 1. **成对方法**

尽管列表式方法已经取得了有希望的性能，但它们仍然存在一些弱点。首先，根据实验结果 [128]，只有基于GPT-4的方法能够取得有竞争力的性能。当使用参数化规模较小的语言模型（例如，具有200亿参数的 FLANUL2）时，列表式方法可能会产生很少的可用的结果，并且性能低于许多监督方法。其次，列表式方法的性能对提示中的文档顺序高度敏感。当文档顺序随机打乱时，列表式方法的性能甚至比 BM25 还差。这些弱点可能是由于现有的流行 LLM 通常没有针对排序任务进行预训练造成的。

一项最新研究 [130] 表明，LLM 本身具有成对文档比较的能力，这比直接输出重新排序的文档列表要简单得多。在本研究中，首先设计了一个成对排序提示（见图 6 (c)），用于比较一个文档是否比另一个文档更相关。然后，设计了几种排序算法，利用这种成对排序提示作为计算单元，对整个文档列表进行重新排序。实验结果表明，使用中等规模的LLM（例如，具有200亿个参数的Flan-UL2），在标准基准测试中取得了最先进的性能，这远小于GPT-4。

* 1. **利用大型语言模型进行训练数据增强**

此外，在重排序领域，研究人员探索了将大型语言模型 (LLM) 用于训练数据增强 [128, 131–133]。例如，ExaRanker [131] 使用 GPT-3.5 为检索数据集生成解释，随后训练一个 seq2seq 排序模型，以针对给定的查询-文档对生成相关性标签及其相应的解释。InPars-Light [132] 被提出作为一种经济高效的方法，通过提示大型语言模型 (LLM) 来合成文档查询。与 InPars-Light [132] 相反，一个新的数据集 ChatGPT-RetrievalQA [133] 是通过基于 LLM 生成对用户查询的响应来构建合成文档而构建的。此外，也有人尝试将 ChatGPT 的文档排序能力提炼到一个专门的模型中 [128]。在这种方法中，首先指示 ChatGPT 直接生成文档排序列表。然后，使用生成的文档列表作为目标，使用各种排序损失（例如，RankNet [123]）训练学生模型（即，DeBERTa-v3-base）。

* 1. **限制**

尽管最近利用大型语言模型 (LLM) 进行文档重排序的研究取得了重大进展，但仍面临一些挑战。例如，考虑到成本和效率，最小化对 LLM API 的调用次数是一个值得研究的问题。此外，虽然现有研究主要集中于将 LLM 应用于开放域数据集（例如 MSMARCO [99]）或基于相关性的文本排序任务，但它们对域内数据集 [112] 和非标准排序数据集 [134] 的适应性仍然是一个需要更全面探索的领域。

1. **阅读器**

随着大型语言模型 (LLM) 在理解、提取和处理文本数据方面展现出令人印象深刻的能力，研究人员探索将信息检索 (IR) 系统的范围从内容排序扩展到答案生成。在这个演变过程中，一个阅读模块被引入到 IR 系统中，以根据文档语料库生成答案。通过集成阅读模块，信息检索系统可以直接向用户呈现结论性段落。与提供文档列表相比，在这种新的范式中，用户可以简单地理解答案段落，而不是分析排名列表。此外，通过根据其生成文本反复向大型语言模型提供文档，最终生成的答案可能比最初检索到的列表更准确、信息更丰富。

* 1. **被动阅读器**

为了生成用户的答案，一种直接的策略是将检索到的文档（根据查询或来自信息检索系统的先前生成的文本）作为输入提供给大型语言模型 (LLM) 以创建段落 [23, 135–144]。通过这种方式，这些方法以独立的方式使用 LLM 和信息检索系统，LLM 充当信息检索系统提供的文档的被动接收者。根据大语言模型 (LLM) 在信息检索系统 (IR) 阅读模块中检索文档的频率，可以将利用 LLM 的策略分为以下三类。

* + 1. **一次检索阅读器**

为了获得有用的参考，使LLMs能够生成对用户查询的响应，一个直观的方法是在开始时根据查询本身检索出排名前列的文档。例如，REALM [135] 采用了这种策略，通过直接将文档内容与原始查询进行关联，根据掩码语言模型预测最终答案。RAG [136] 遵循这种策略，但应用了生成式语言模型范式。然而，这两种方法只使用参数有限的语言模型，例如 BERT 和 BART。最近的方法，例如 REPLUG [137] 和 Atlas [138]，通过利用 GPT 和 T5 等大型语言模型来生成响应，对它们进行了改进。为了获得更好的答案生成性能，这些模型通常在问答任务上微调大型语言模型。然而，由于计算资源有限，许多方法 [139, 140] 选择使用提示的方式来引导大型语言模型进行生成，因为这样可以利用更大的模型。为了更好地促进固定LLM执行阅读任务，RETA-LLM [141] 将整个复杂的生成任务分解为阅读管道中的几个简单模块。这些模块包括一个用于细化查询意图的查询重写模块，一个用于将参考长度与LLM限制对齐的段落提取模块，以及一个用于确认生成答案中不存在虚假信息的真实性验证模块。

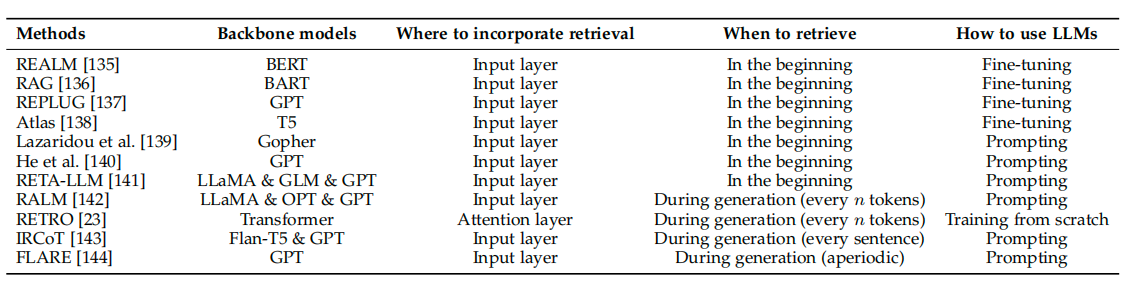
* + 1. **周期性检索阅读器**

然而，在生成长篇结论性答案时，研究表明 [23, 142] 仅使用原始用户意图检索到的参考文献，如一次检索阅读器，可能是不够的。例如，在提供关于“巴拉克·奥巴马”的段落时，语言模型可能需要关于他大学的额外知识，而这些知识可能不包含在简单搜索初始查询的结果中。总之，语言模型可能需要额外的参考文献来支持生成过程中的后续生成，这可能需要多次检索过程。为了解决这个问题，出现了诸如 RETRO [23] 和 RALM [142] 之类的解决方案，强调基于原始查询和同时生成的文本定期收集文档（每生成 n个标记触发检索）。这样，在生成关于巴拉克·奥巴马大学生涯的文本时，LLM 可以接收额外的文档作为补充材料。对额外参考文献的需求突出了多次检索迭代以确保后续答案生成稳健性的必要性。值得注意的是，RETRO [23] 引入了一种新方法，在 Transformer 注意力计算中，将生成文本与参考文献之间的交叉注意力结合起来，而不是直接将参考文献嵌入到 LLM 的输入文本中。由于 RETRO 在 Transformer 结构中引入了额外的交叉注意力模块，因此它从头开始训练该模型。然而，这两种方法主要依赖于连续的 n 标记来分离生成和检索文档，这可能在语义上不连续，并可能导致收集的参考资料存在噪声和无用信息。为了解决这个问题，一些方法，例如 IRCoT [143]，也探索了为每个生成的句子检索文档，这是一种更完整的语义结构。

* + 1. **非周期性检索阅读器**

在上述策略中，检索系统以周期性方式向LLM提供文档。然而，强制性频率检索文档可能会导致检索时间不匹配，并且成本高昂。最近，FLARE [144] 通过根据生成文本的概率自动确定检索时间来解决这个问题。由于概率可以作为文本生成过程中大型语言模型置信度的指标 [145, 146]，生成的词语的低概率可能表明大型语言模型需要额外的知识。具体来说，当某个词语的概率低于预定义的阈值时，FLARE 会利用信息检索系统根据正在生成的句子检索参考资料，同时移除这些低概率词语。FLARE 采用这种策略，仅基于生成词语的概率来提示大型语言模型生成答案，避免了微调的需要，同时仍然保持有效性。

表 5. 具有被动阅读模块的现有方法比较。REALM 和 RAG 不使用大型语言模型，但其框架已广泛应用于许多后续方法。



我们将这些被动读者方法总结在表 5 中，考虑了各种方面，例如主干语言模型、检索参考的插入点、使用检索模型的时间以及用于 LLM 的调优策略。

* 1. **主动阅读器**

然而，基于被动读者的方法将信息检索系统和生成式语言模型分开。这意味着大型语言模型只能被动地利用信息检索系统提供的参考，而无法像人类互动那样与信息检索系统进行交互，例如发出查询以寻求信息。

为了使LLMs能够主动使用搜索引擎，SelfAsk [147] 和 DSP [148] 尝试使用少样本提示来触发LLMs，当它们认为需要时，触发它们进行搜索查询。例如，在一个查询为“现存最高的木桁架塔何时建造？”的场景中，这些被提示的LLMs可以决定搜索查询“现存最高的木桁架塔是什么”，以收集必要的参考资料，因为它们发现该查询无法直接回答。一旦获取了关于塔的信息，他们就可以迭代地查询信息检索系统以获取更多细节，直到他们决定生成最终答案而不是提出问题。值得注意的是，这些方法涉及使用信息检索系统为大型语言模型构建单个推理链。MRC [149] 通过提示大型语言模型探索多个推理链，并随后使用大型语言模型将所有生成的答案组合在一起，进一步改进了这些方法。

与提示大型语言模型不同，WebGPT [24] 采取了一种替代方法，通过训练 GPT-3 模型自动使用搜索引擎。这是通过应用强化学习框架实现的，在这个框架中，为 GPT-3 模型构建了一个模拟环境。具体来说，WebGPT 模型使用特殊标记来执行操作，例如查询、滚动浏览排名和引用搜索引擎上的参考资料。这种创新方法允许 GPT-3 模型使用搜索引擎进行文本生成，从而提高了生成文本的可靠性和实时性。一些后续模型 [150] 将这种范式扩展到了中文问答领域。

* 1. **限制**

一些应用检索增强生成策略的 IR 系统，例如新必应和 Langchain，已经投入商业使用。然而，这种新颖的检索增强内容生成系统也面临一些挑战。这些挑战包括有效的查询重构、最佳检索频率、正确的文档理解、准确的段落提取以及有效的内容摘要。解决这些挑战对于在实际应用中充分发挥这种方法的潜力至关重要。

1. **未来方向**

在本调查中，我们对增强了 LLM 的 IR 系统的最新进展进行了全面回顾。由于将 LLM 集成到 IR 系统中仍处于早期阶段，因此仍存在许多机遇和挑战。在本节中，我们将讨论一些潜在的未来方向。未来的方向是根据最近一篇观点论文 [50] 中提出的分类法进行组织的。

* 1. **索引**

正如第 4.2.2 节所述，LLM 的出现为生成式检索器生成文档标识符以用于检索任务铺平了道路。通过这种方式，文档索引或关于文档的知识被编码在 LLM 的参数中。然而，LLM 参数固有的静态性质，加上昂贵的微调过程，给生成式检索器更新新文档的文档索引带来了挑战。因此，需要探索构建生成式检索增量索引的方法。

此外，网络上的内容可以是多模态的，涵盖了不同用户的需求，包括文本、图像、音频和视频。然而，以前的工作主要利用候选者的语义信息作为文档标识符，并专注于基于文本的检索（单模态）。随着多模态LLMs的不断发展，为有效的多模态检索设计和启用LLMs成为一项值得深入研究的挑战。

* 1. **用户建模**

信息检索中的用户建模是指理解用户查询、偏好和搜索行为以增强搜索体验的过程。由于大型语言模型 (LLM) 在理解自然语言方面具有出色的能力，将 LLM 整合到用户建模中为改进个性化搜索和交互提供了巨大潜力。几个有希望的方向被强调：

1. 增强查询理解。用户查询往往模棱两可或多方面，阻碍搜索引擎理解用户的真实搜索意图。通过分析查询文本，LLMs可以识别查询中的消歧词或潜在的子意图[151]，这有助于澄清用户的需求并更准确地推断用户意图。
2. 改善个性化搜索。LLMs 可以通过分析用户特定信息来协助个性化搜索。通过整合用户的搜索历史（例如，发出的查询、点击行为和停留时间），LLMs 可以构建准确的用户画像，并相应地重新排序搜索结果，提供更好的结果。
3. 推进对话式搜索界面。大型语言模型 (LLM) 在改进对话式搜索界面方面可以发挥关键作用，提供更自然、更直观的交互体验。通过整合历史交互信息，LLM 可以根据用户偏好调整系统响应，提供更个性化、更有效的对话体验。

通过利用大型语言模型的潜在能力，这些方向为信息检索中更复杂和定制的用户建模铺平了道路。随着大型语言模型的不断发展，它们彻底改变以用户为中心的搜索范式的潜力仍然很重要。

* 1. **匹配或查询**

利用大型语言模型 (LLM) 提升信息检索 (IR) 中的排序模型已受到广泛关注，它承诺能够更好地理解查询和文档，从而提高排序性能。然而，尽管该领域取得了进展，但仍存在一些需要解决的局限性和挑战：

1. 缓解 LLM 延迟。LLM 由于其庞大的参数和世界知识，在在线应用中往往会产生高延迟。鉴于搜索引擎需要及时响应，平衡基于 LLM 的排序模型的效率和及时生成排序列表成为一项重要挑战。
2. 为排序任务定制大型语言模型。现有研究表明，大型语言模型能够在零样本或少样本场景下泛化到各种下游任务。然而，少样本大型语言模型的性能仍然不如在特定任务训练数据上微调的模型 [105, 107]，因为它们缺乏特定任务的知识。因此，有必要对大型语言模型进行排序任务的微调，以提高其领域内能力。此外，由于资源限制，对参数超过 100 亿的大型语言模型进行微调仍然未得到充分探索。现有方法主要集中在参数相对较小的模型 [110]，例如 T5。然而，许多研究证实，更大的参数可以带来指数级的性能提升 [113, 119]。因此，如何对具有数百亿或数千亿参数的大型语言模型进行微调（或从头开始预训练）以用于排序任务仍然是一个开放且紧迫的问题。
3. 信息检索框架的演变。除了将大型语言模型用作排序模型之外，大型语言模型的出现带来的一个更重要的问题是：在大型语言模型时代，传统的 IR 框架是否仍然必要？传统的 IR 旨在返回与发布的查询相关的文档排序列表。然而，生成式语言模型的发展引入了一种新的范式：大型语言模型直接生成对输入问题的答案。显然，单个大型语言模型难以确保答案的及时性、可靠性和真实性。因此，使用 IR 模型为大型语言模型提供可靠的参考是必不可少的，这导致了两个进一步的问题：什么是好的参考对于大型语言模型生成答案，以及如何利用信息检索模型提供的参考文献来生成更好的答案。
   1. **评估**

由于其强大的上下文理解和文本生成能力，LLMs 在信息检索领域引起了广泛关注。为了验证基于 LLM 的信息检索方法的有效性，开发合适的指标和数据集至关重要。传统的评估指标主要集中在比较信息检索模型的检索结果与真实结果，例如精确率、召回率、平均倒数排名 (MRR) [152]、平均精确率 (MAP) 和归一化折现累积增益 (nDCG) [153]。然而，LLM 在信息检索中的广泛应用显著提升了其生成能力。因此，迫切需要新的评估策略来有效评估基于 LLM 的模型在信息检索中的性能。以往的文本生成评估指标存在一些局限性，包括：（1）依赖词汇匹配：像 BLEU [154] 这样的方法主要基于 n-gram 匹配来评估生成输出的质量。这种方法没有考虑词汇多样性和语境语义。因此，模型可能倾向于生成常见的短语或句子结构，而不是生成有创意和新颖的内容。(2) 对细微差异的敏感度不足：传统的评估方法可能对生成的输出中的细微差异不敏感。例如，如果生成的输出与参考答案在语义上存在细微差异，但其他方面相似，传统方法可能无法捕捉到这些细微的差别。如何评估生成内容的质量仍然是一个挑战。

* 1. **交互**

随着大型语言模型 (LLM) 能力的不断提升，用户与信息检索 (IR) 系统之间的未来交互方式将发生重大改变。由于 LLM 强大的自然语言处理和理解能力，传统的基于关键词的搜索范式预计将逐步被对话式搜索所取代。大型语言模型将促进全文档理解，使用户能够提出复杂、上下文感知的问题并获得全面的答案。通过使用外部插件，大型语言模型有可能理解图像和语音输入，从而促进复杂的跨模态和跨模态检索。此外，大型语言模型可以根据用户的历史搜索行为和偏好主动提供信息和推荐，从而增强整体搜索体验。这些应用展示了大型语言模型在革新现有信息检索系统交互范式的潜力，从而在未来为用户提供更个性化和无缝的体验。

1. **结论**

在本调查中，我们对大型语言模型 (LLM) 对信息检索 (IR) 的变革性影响进行了全面的探索，涵盖了各个方面。我们根据其功能将现有方法划分为不同的类别：查询重写、检索、重排序和阅读器模块。在查询重写领域，LLM 在理解模糊或多方面查询方面展现出其有效性，提高了意图识别的准确性。在检索方面，LLM 通过实现查询和文档之间更细致的匹配，同时考虑上下文，提高了检索准确性。在重排序领域，增强了大型语言模型 (LLM) 的模型在重新排序结果时考虑了更细粒度的语言细微差别。在信息检索 (IR) 系统中加入阅读器模块代表着朝着生成全面响应而不是仅仅提供文档列表迈出的重要一步。将大型语言模型 (LLM) 整合到信息检索 (IR) 系统中，从根本上改变了用户与信息和知识互动的方式。从查询重写到检索、重排序和阅读模块，LLMs 已经通过先进的语言理解、语义表示和上下文敏感处理，丰富了信息检索流程的各个方面。随着该领域不断发展，LLMs 在信息检索中的旅程预示着未来将出现更加个性化、精确和以用户为中心的搜索体验。

**参考文献**

1. Y. Wu, W. Wu, C. Xing, M. Zhou, and Z. Li, “Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots,” in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, R. Barzilay and M. Kan, Eds. Association for Computational Linguistics, 2017, pp. 496–505.
2. H. Shum, X. He, and D. Li, “From eliza to xiaoice: challenges and opportunities with social chatbots,” Frontiers Inf. Technol. Electron. Eng., vol. 19, no. 1, pp. 10–26, 2018.
3. V. Karpukhin, B. Oguz, S. Min, P. S. H. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense passage retrieval for open-domain question answering,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 6769–6781.
4. R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Comput. Surv., vol. 40, no. 2, pp. 5:1–5:60, 2008.
5. C. Yuan, W. Zhou, M. Li, S. Lv, F. Zhu, J. Han, and S. Hu, “Multi-hop selector network for multiturn response selection in retrieval-based chatbots,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 37, 2019, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, 2019, pp. 111–120.
6. Y. Zhu, J. Nie, K. Zhou, P. Du, and Z. Dou, “Content selection network for document-grounded retrievalbased chatbots,” in Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I, ser. Lecture Notes in Computer Science, D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, Eds., vol. 12656. Springer, 2021, pp. 755–769.
7. Y. Zhu, J. Nie, K. Zhou, P. Du, H. Jiang, and Z. Dou, “Proactive retrieval-based chatbots based on relevant knowledge and goals,” in SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, Eds. ACM, 2021, pp. 2000– 2004.
8. H. Qian, Z. Dou, Y. Zhu, Y. Ma, and J. Wen, “Learning implicit user profiles for personalized retrieval-based chatbot,” CoRR, vol. abs/2108.07935, 2021.
9. Y. Qu, Y. Ding, J. Liu, K. Liu, R. Ren, W. X. Zhao, D. Dong, H. Wu, and H. Wang, “Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 611, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-T¨ur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021, pp. 5835–5847.
10. Y. Arens, C. A. Knoblock, and W. Shen, “Query reformulation for dynamic information integration,” J. Intell. Inf. Syst., vol. 6, no. 2/3, pp. 99–130, 1996.
11. J. Huang and E. N. Efthimiadis, “Analyzing and evaluating query reformulation strategies in web search logs,” in Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009, D. W. Cheung, I. Song, W. W. Chu, X. Hu, and J. Lin, Eds. ACM, 2009, pp. 77–86.
12. R. F. Nogueira, W. Yang, K. Cho, and J. Lin, “Multistage document ranking with BERT,” CoRR, vol. abs/1910.14424, 2019.
13. R. F. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, “Document ranking with a pretrained sequence-to-sequence model,” in EMNLP (Findings), ser. Findings of ACL, vol. EMNLP 2020. Association for Computational Linguistics, 2020, pp. 708–718.
14. Y. Zhu, J. Nie, Z. Dou, Z. Ma, X. Zhang, P. Du, X. Zuo, and H. Jiang, “Contrastive learning of user behavior sequence for context-aware document ranking,” in CIKM ’21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, and H. Tong, Eds. ACM, 2021, pp. 2780–2791.
15. J. Teevan, S. T. Dumais, and E. Horvitz, “Personalizing search via automated analysis of interests and activities,” in SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005, R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, Eds. ACM, 2005, pp. 449–456.
16. P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui, “Modeling the impact of short- and long-term behavior on search personalization,” in The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR ’12, Portland, OR, USA, August 12-16, 2012, W. R. Hersh, J. Callan, Y. Maarek, and M. Sanderson, Eds. ACM, 2012, pp. 185–194.
17. S. Ge, Z. Dou, Z. Jiang, J. Nie, and J. Wen, “Personalizing search results using hierarchical RNN with query-aware attention,” in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, A. Cuzzocrea, J. Allan, N. W. Paton, D.
18. Y. Zhou, Z. Dou, Y. Zhu, and J. Wen, “PSSL: selfsupervised learning for personalized search with contrastive sampling,” in CIKM ’21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, and H. Tong, Eds. ACM, 2021, pp. 2749– 2758.
19. J. G. Carbonell and J. Goldstein, “The use of mmr, diversity-based reranking for reordering documents and producing summaries,” in SIGIR ’98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia, W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, Eds. ACM, 1998, pp. 335–336.
20. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, “Diversifying search results,” in Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009, R. Baeza-Yates, P. Boldi, B. A. Ribeiro-Neto, and B. B. Cambazoglu, Eds. ACM, 2009, pp. 5–14.
21. J. Liu, Z. Dou, X. Wang, S. Lu, and J. Wen, “DVGAN: A minimax game for search result diversiﬁcation combining explicit and implicit features,” in Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 479–488.
22. Z. Su, Z. Dou, Y. Zhu, X. Qin, and J. Wen, “Modeling intent graph for search result diversiﬁcation,” in SIGIR ’21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai, Eds. ACM, 2021, pp. 736–746.
23. S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, and L. Sifre, “Improving language models by retrieving from trillions of tokens,” in International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari,´ G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 2206–2240.
24. R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, X. Jiang, K. Cobbe, T. Eloundou, G. Krueger, K. Button, M. Knight, B. Chess, and J. Schulman, “Webgpt: Browser-assisted question-answering with human feedback,” CoRR, vol. abs/2112.09332, 2021.
25. G. Salton and M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill Book Company, 1984.
26. G. Salton, A. Wong, and C. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.
27. F. Song and W. B. Croft, “A general language model for information retrieval,” in Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management, Kansas City, Missouri, USA, November 2-6, 1999. ACM, 1999, pp. 316–321.
28. J. Martineau and T. Finin, “Delta TFIDF: an improved feature space for sentiment analysis,” in Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, May 17- 20, 2009, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. The AAAI Press, 2009.
29. S. E. Robertson, S. Walker, S. Jones, M. HancockBeaulieu, and M. Gatford, “Okapi at TREC-3,” in Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, ser. NIST Special Publication, D. K. Harman, Ed., vol. 500-225. National Institute of Standards and Technology (NIST), 1994, pp. 109–126.
30. J. Guo, Y. Fan, Q. Ai, and W. B. Croft, “A deep relevance matching model for ad-hoc retrieval,” in Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016, S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, and P. Sondhi, Eds. ACM, 2016, pp. 55–64.
31. L. Xiong, C. Xiong, Y. Li, K. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk, “Approximate nearest neighbor negative contrastive learning for dense text retrieval,” in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
32. J. Lin, R. F. Nogueira, and A. Yates, Pretrained Transformers for Text Ranking: BERT and Beyond, ser. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021.
33. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
34. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.
35. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozier` e, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efﬁcient foundation language models,” CoRR, vol. abs/2302.13971, 2023.
36. J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J. Wen, “Recommendation as instruction following: A large language model empowered recommendation approach,” CoRR, vol. abs/2305.07001, 2023.
37. Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. J. McAuley, and W. X. Zhao, “Large language models are zeroshot rankers for recommender systems,” CoRR, vol. abs/2305.08845, 2023.
38. Y. Xi, W. Liu, J. Lin, J. Zhu, B. Chen, R. Tang, W. Zhang, R. Zhang, and Y. Yu, “Towards open-world recommendation with knowledge augmentation from large language models,” CoRR, vol. abs/2306.10933, 2023.
39. W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, and Q. Li, “Recommender systems in the era of large language models (llms),” CoRR, vol. abs/2307.02046, 2023.
40. J. Li, Y. Liu, W. Fan, X. Wei, H. Liu, J. Tang, and Q. Li, “Empowering molecule discovery for moleculecaption translation with large language models: A chatgpt perspective,” CoRR, vol. abs/2306.06615, 2023.
41. J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, “Emergent abilities of large language models,” Trans. Mach. Learn. Res., vol. 2022, 2022.
42. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, “Chain-ofthought prompting elicits reasoning in large language models,” in NeurIPS, 2022.
43. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” ACM Comput. Surv., vol. 55, no. 9, pp. 195:1–195:35, 2023.
44. X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” CoRR, vol. abs/2003.08271, 2020.
45. Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, “A comprehensive survey of ai-generated content (AIGC): A history of generative AI from GAN to chatgpt,” CoRR, vol. abs/2303.04226, 2023.
46. J. Li, T. Tang, W. X. Zhao, and J. Wen, “Pretrained language model for text generation: A survey,” in Proceedings of the Thirtieth International Joint Conference on Artiﬁcial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, Z. Zhou, Ed. ijcai.org, 2021, pp. 4492–4499.
47. Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, L. Li, and Z. Sui, “A survey for in-context learning,” CoRR, vol. abs/2301.00234, 2023.
48. J. Huang and K. C. Chang, “Towards reasoning in large language models: A survey,” in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. BoydGraber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 1049–1065.
49. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen, “A survey of large language models,” CoRR, vol. abs/2303.18223, 2023.
50. Q. Ai, T. Bai, Z. Cao, Y. Chang, J. Chen, Z. Chen, Z. Cheng, S. Dong, Z. Dou, F. Feng, S. Gao, J. Guo, X. He, Y. Lan, C. Li, Y. Liu, Z. Lyu, W. Ma, J. Ma, Z. Ren, P. Ren, Z. Wang, M. Wang, J. Wen, L. Wu, X. Xin, J. Xu, D. Yin, P. Zhang, F. Zhang, W. Zhang, M. Zhang, and X. Zhu, “Information retrieval meets large language models: A strategic report from chinese IR community,” CoRR, vol. abs/2307.09751, 2023.
51. X. Liu and W. B. Croft, “Statistical language modeling for information retrieval,” Annu. Rev. Inf. Sci. Technol., vol. 39, no. 1, pp. 1–31, 2005.
52. B. Mitra and N. Craswell, “Neural models for information retrieval,” CoRR, vol. abs/1705.01509, 2017.
53. W. X. Zhao, J. Liu, R. Ren, and J. Wen, “Dense text retrieval based on pretrained language models: A survey,” CoRR, vol. abs/2211.14876, 2022.
54. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a uniﬁed text-totext transformer,” J. Mach. Learn. Res., vol. 21, pp. 140:1–140:67, 2020.
55. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 2227–2237.
56. J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186.
57. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998– 6008.
58. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7871–7880.
59. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” CoRR, vol. abs/2001.08361, 2020.
60. A. Clark, D. de Las Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. A. Hechtman, T. Cai, S. Borgeaud, G. van den Driessche, E. Rutherford, T. Hennigan, M. J. Johnson, A. Cassirer, C. Jones, E. Buchatskaya, D. Budden, L. Sifre, S. Osindero, O. Vinyals, M. Ranzato, J. W. Rae, E. Elsen, K. Kavukcuoglu, and K. Simonyan, “Uniﬁed scaling laws for routed language models,” in International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari,´ G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 4057–4086.
61. L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, “Uniﬁed language model pre-training for natural language understanding and generation,” in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alche-´ Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13 042– 13 054.
62. L. Xue, N. Constant, A. Roberts, M. Kale, R. AlRfou, A. Siddhant, A. Barua, and C. Raffel, “mt5: A massively multilingual pre-trained text-to-text transformer,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur¨ , I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Association for Computational Linguistics, 2021, pp. 483–498.
63. V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chafﬁn, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry´ , J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, and A. M. Rush, “Multitask prompted training enables zero-shot task generalization,” in The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
64. H. Bao, L. Dong, F. Wei, W. Wang, N. Yang, X. Liu, Y. Wang, J. Gao, S. Piao, M. Zhou, and H. Hon, “Unilmv2: Pseudo-masked language models for uniﬁed language model pre-training,” in Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 642–652.
65. A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, and J. Tang, “GLM-130B: an open bilingual pre-trained model,” in The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.
66. W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efﬁcient sparsity,” J. Mach. Learn. Res., vol. 23, pp. 120:1–120:39, 2022.
67. Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alche-Buc,´ E. B. Fox, and R. Garnett, Eds., 2019, pp. 5754–5764.
68. S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. S. Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, “Gptneox-20b: An open-source autoregressive language model,” CoRR, vol. abs/2204.06745, 2022.
69. J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, H. F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, E. Rutherford, T. Hennigan, J. Menick, A. Cassirer, R. Powell, G. van den Driessche, L. A. Hendricks, M. Rauh, P. Huang, A. Glaese, J. Welbl, S. Dathathri, S. Huang, J. Uesato, J. Mellor, I. Higgins, A. Creswell, N. McAleese, A. Wu, E. Elsen, S. M. Jayakumar, E. Buchatskaya, D. Budden, E. Sutherland, K. Simonyan, M. Paganini, L. Sifre, L. Martens, X. L. Li, A. Kuncoro, A. Nematzadeh, E. Gribovskaya, D. Donato, A. Lazaridou, A. Mensch, J. Lespiau, M. Tsimpoukelli, N. Grigorev, D. Fritz, T. Sottiaux, M. Pajarskas, T. Pohlen, Z. Gong, D. Toyama, C. de Masson d’Autume, Y. Li, T. Terzi, V. Mikulik, I. Babuschkin, A. Clark, D. de Las Casas, A. Guy, C. Jones, J. Bradbury, M. J. Johnson, B. A. Hechtman, L. Weidinger, I. Gabriel, W. Isaac, E. Lockhart, S. Osindero, L. Rimell, C. Dyer, O. Vinyals, K. Ayoub, J. Stanway, L. Bennett, D. Hassabis, K. Kavukcuoglu, and G. Irving, “Scaling language models: Methods, analysis & insights from training gopher,” CoRR, vol. abs/2112.11446, 2021.
70. N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. P. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. S. MeierHellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, “Glam: Efﬁcient scaling of language models with mixture-of-experts,” in International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari,´ G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 5547–5569.
71. Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang, “ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation,” CoRR, vol. abs/2107.02137, 2021.
72. S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, “OPT: open pre-trained transformer language models,” CoRR, vol. abs/2205.01068, 2022.
73. R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, Y. Li, H. Lee, H. S. Zheng, A. Ghafouri, M. Menegali, Y. Huang, M. Krikun, D. Lepikhin, J. Qin, D. Chen, Y. Xu, Z. Chen, A. Roberts, M. Bosma, Y. Zhou, C. Chang, I. Krivokon, W. Rusch, M. Pickett, K. S. Meier-Hellstern, M. R. Morris, T. Doshi, R. D. Santos, T. Duke, J. Soraker, B. Zevenbergen, V. Prabhakaran, M. Diaz, B. Hutchinson, K. Olson, A. Molina, E. Hoffman-John, J. Lee, L. Aroyo, R. Rajakumar, A. Butryna, M. Lamm, V. Kuzmina, J. Fenton, A. Cohen, R. Bernstein, R. Kurzweil, B. A. y Arcas, C. Cui, M. Croak, E. H. Chi, and Q. Le, “Lamda: Language models for dialog applications,” CoRR, vol. abs/2201.08239, 2022.
74. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. MeierHellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, “Palm: Scaling language modeling with pathways,” CoRR, vol. abs/2204.02311, 2022.
75. T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilic, D. Hesslow, R. Castagne,´ A. S. Luccioni, F. Yvon, M. Galle,´ J. Tow, A. M. Rush, S. Biderman, A. Webson, P. S. Ammanamanchi, T. Wang, B. Sagot, N. Muennighoff, A. V. del Moral, O. Ruwase, R. Bawden, S. Bekman, A. McMillan-Major, I. Beltagy, H. Nguyen, L. Saulnier, S. Tan, P. O. Suarez, V. Sanh, H. Laurenc¸on, Y. Jernite, J. Launay, M. Mitchell, C. Raffel, A. Gokaslan, A. Simhi, A. Soroa, A. F. Aji, A. Alfassy, A. Rogers, A. K. Nitzav, C. Xu, C. Mou, C. Emezue, C. Klamm, C. Leong, D. van Strien, D. I. Adelani, and et al., “BLOOM: A 176b-parameter open-access multilingual language model,” CoRR, vol. abs/2211.05100, 2022.
76. A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo, Y. Wu, B. Neyshabur, G. Gur-Ari, and V. Misra, “Solving quantitative reasoning problems with language models,” in NeurIPS, 2022.
77. OpenAI, “GPT-4 technical report,” CoRR, vol. abs/2303.08774, 2023.
78. J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, “Training compute-optimal large language models,” CoRR, vol. abs/2203.15556, 2022.
79. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
80. X. L. Li and P. Liang, “Preﬁx-tuning: Optimizing continuous prompts for generation,” in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1- 6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 4582–4597.
81. B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efﬁcient prompt tuning,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 3045–3059.
82. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efﬁcient ﬁnetuning of quantized llms,” CoRR, vol. abs/2305.14314, 2023.
83. N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, and C. Wade, “Umass at TREC 2004: Novelty and HARD,” in Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004, ser. NIST Special Publication, E. M. Voorhees and L. P. Buckland, Eds., vol. 500-261. National Institute of Standards and Technology (NIST), 2004.
84. D. Metzler and W. B. Croft, “Latent concept expansion using markov random ﬁelds,” in SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007, W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, Eds. ACM, 2007, pp. 311–318.
85. C. Zhai and J. D. Lafferty, “Model-based feedback in the language modeling approach to information retrieval,” in Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001. ACM, 2001, pp. 403–410.
86. D. Metzler and W. B. Croft, “A markov random ﬁeld model for term dependencies,” in SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August 15-19, 2005, R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, and J. Tait, Eds. ACM, 2005, pp. 472–479.
87. L. Wang, N. Yang, and F. Wei, “Query2doc: Query expansion with large language models,” CoRR, vol. abs/2303.07678, 2023.
88. K. Mao, Z. Dou, H. Chen, F. Mo, and H. Qian, “Large language models know your contextual search intent: A prompting framework for conversational search,” CoRR, vol. abs/2303.06573, 2023.
89. L. Gao, X. Ma, J. Lin, and J. Callan, “Precise zero-shot dense retrieval without relevance labels,” CoRR, vol. abs/2212.10496, 2022.
90. R. Jagerman, H. Zhuang, Z. Qin, X. Wang, and M. Bendersky, “Query expansion by prompting large language models,” CoRR, vol. abs/2305.03653, 2023.
91. X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, “Query rewriting for retrieval-augmented large language models,” CoRR, vol. abs/2305.14283, 2023.
92. I. Mackie, I. Sekulic, S. Chatterjee, J. Dalton, and F. Crestani, “GRM: generative relevance modeling using relevance-aware sample estimation for document retrieval,” CoRR, vol. abs/2306.09938, 2023.
93. K. Srinivasan, K. Raman, A. Samanta, L. Liao, L. Bertelli, and M. Bendersky, “QUILL: query intent with large language models using retrieval augmentation and multi-stage distillation,” in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022, Y. Li and A. Lazaridou, Eds. Association for Computational Linguistics, 2022, pp. 492–501.
94. J. Feng, C. Tao, X. Geng, T. Shen, C. Xu, G. Long, D. Zhao, and D. Jiang, “Knowledge reﬁnement via interaction between search engines and large language models,” CoRR, vol. abs/2305.07402, 2023.
95. I. Mackie, S. Chatterjee, and J. Dalton, “Generative and pseudo-relevant feedback for sparse, dense and learned sparse retrieval,” CoRR, vol. abs/2305.07477, 2023.
96. T. Shen, G. Long, X. Geng, C. Tao, T. Zhou, and D. Jiang, “Large language models are strong zero-shot retriever,” CoRR, vol. abs/2304.14233, 2023.
97. M. Alaoﬁ, L. Gallagher, M. Sanderson, F. Scholer, and P. Thomas, “Can generative llms create query variants for test collections? an exploratory study,” in Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, and B. Poblete, Eds. ACM, 2023, pp. 1869–1873.
98. W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang, “Generate rather than retrieve: Large language models are strong context generators,” in The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023.
99. T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, “MS MARCO: A human generated machine reading comprehension dataset,” in CoCo@NIPS, ser. CEUR Workshop Proceedings, vol. 1773. CEUR-WS.org, 2016.
100. T. Kwiatkowski, J. Palomaki, O. Redﬁeld, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov, “Natural questions: a benchmark for question answering research,” Trans. Assoc. Comput. Linguistics, vol. 7, pp. 452–466, 2019.
101. G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021, pp. 874–880.
102. J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: a large-scale dataset for fact extraction and veriﬁcation,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), M. A. Walker, H. Ji, and A. Stent, Eds. Association for Computational Linguistics, 2018, pp. 809–819.
103. D. Alexander, W. Kusa, and A. P. de Vries, “ORCASI: queries annotated with intent using weak supervision,” in SIGIR ’22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, E. Amigo,´ P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds. ACM, 2022, pp. 3057–3066.
104. I. Mackie, I. Sekulic, S. Chatterjee, J. Dalton, and F. Crestani, “GRM: generative relevance modeling using relevance-aware sample estimation for document retrieval,” CoRR, vol. abs/2306.09938, 2023.
105. L. H. Bonifacio, H. Abonizio, M. Fadaee, and R. F. Nogueira, “Inpars: Data augmentation for information retrieval using large language models,” CoRR, vol. abs/2202.05144, 2022.
106. V. Jeronymo, L. H. Bonifacio, H. Abonizio, M. Fadaee, R. de Alencar Lotufo, J. Zavrel, and R. F. Nogueira, “Inpars-v2: Large language models as efﬁcient dataset generators for information retrieval,” CoRR, vol. abs/2301.01820, 2023.
107. Z. Dai, V. Y. Zhao, J. Ma, Y. Luan, J. Ni, J. Lu, A. Bakalov, K. Guu, K. B. Hall, and M. Chang, “Promptagator: Few-shot dense retrieval from 8 examples,” in ICLR. OpenReview.net, 2023.
108. R. Meng, Y. Liu, S. Yavuz, D. Agarwal, L. Tu, N. Yu, J. Zhang, M. Bhat, and Y. Zhou, “Augtriever: Unsupervised dense retrieval by scalable data augmentation,” 2023.
109. J. Saad-Falcon, O. Khattab, K. Santhanam, R. Florian, M. Franz, S. Roukos, A. Sil, M. A. Sultan, and C. Potts, “UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers,” CoRR, vol. abs/2303.00807, 2023.
110. Z. Peng, X. Wu, and Y. Fang, “Soft prompt tuning for augmenting dense retrieval with large language models,” 2023.
111. D. S. Sachan, M. Lewis, D. Yogatama, L. Zettlemoyer, J. Pineau, and M. Zaheer, “Questions are all you need to train a dense passage retriever,” Transactions of the Association for Computational Linguistics, vol. 11, pp. 600–616, 2023.
112. N. Thakur, N. Reimers, A. Ruckl¨ e,´ A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” in NeurIPS Datasets and Benchmarks, 2021.
113. A. Neelakantan, T. Xu, R. Puri, A. Radford, J. M. Han, J. Tworek, Q. Yuan, N. Tezak, J. W. Kim, C. Hallacy, J. Heidecke, P. Shyam, B. Power, T. E. Nekoul, G. Sastry, G. Krueger, D. Schnurr, F. P. Such, K. Hsu, M. Thompson, T. Khan, T. Sherbakov, J. Jang, P. Welinder, and L. Weng, “Text and code embeddings by contrastive pre-training,” CoRR, vol. abs/2201.10005, 2022.
114. J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Abr´ ego, J. Ma, V. Y. Zhao, Y. Luan, K. B. Hall, M. Chang, and Y. Yang, “Large dual encoders are generalizable retrievers,” in EMNLP. Association for Computational Linguistics, 2022, pp. 9844–9855.
115. A. Asai, T. Schick, P. S. H. Lewis, X. Chen, G. Izacard, S. Riedel, H. Hajishirzi, and W. Yih, “Task-aware retrieval with instructions,” in ACL (Findings). Association for Computational Linguistics, 2023, pp. 3650– 3675.
116. D. Metzler, Y. Tay, D. Bahri, and M. Najork, “Rethinking search: making domain experts out of dilettantes,” SIGIR Forum, vol. 55, no. 1, pp. 13:1–13:27, 2021.
117. Y. Zhou, J. Yao, Z. Dou, L. Wu, and J. Wen, “Dynamicretriever: A pre-trained model-based IR system without an explicit index,” Mach. Intell. Res., vol. 20, no. 2, pp. 276–288, 2023.
118. J. Chen, R. Zhang, J. Guo, Y. Liu, Y. Fan, and X. Cheng, “Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks,” in Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022, M. A. Hasan and L. Xiong, Eds. ACM, 2022, pp. 191–200.
119. Y. Tay, V. Tran, M. Dehghani, J. Ni, D. Bahri, H. Mehta, Z. Qin, K. Hui, Z. Zhao, J. P. Gupta, T. Schuster, W. W. Cohen, and D. Metzler, “Transformer memory as a differentiable search index,” in NeurIPS, 2022.
120. N. Ziems, W. Yu, Z. Zhang, and M. Jiang, “Large language models are built-in autoregressive search engines,” in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 2666–2678.
121. J. Ju, J. Yang, and C. Wang, “Text-to-text multi-view learning for passage re-ranking,” in SIGIR. ACM, 2021, pp. 1803–1807.
122. R. Pradeep, R. F. Nogueira, and J. Lin, “The expandomono-duo design pattern for text ranking with pretrained sequence-to-sequence models,” CoRR, vol. abs/2101.05667, 2021.
123. C. J. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. N. Hullender, “Learning to rank using gradient descent,” in ICML, ser. ACM International Conference Proceeding Series, vol. 119. ACM, 2005, pp. 89–96.
124. H. Zhuang, Z. Qin, R. Jagerman, K. Hui, J. Ma, J. Lu, J. Ni, X. Wang, and M. Bendersky, “Rankt5: Finetuning T5 for text ranking with ranking losses,” CoRR, vol. abs/2210.10634, 2022.
125. P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Re,´ D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. J. Orr, L. Zheng, M. Yuksekg¨ on¨ ul,¨ M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, “Holistic evaluation of language models,” CoRR, vol. abs/2211.09110, 2022.
126. D. S. Sachan, M. Lewis, M. Joshi, A. Aghajanyan, W. Yih, J. Pineau, and L. Zettlemoyer, “Improving passage retrieval with zero-shot question generation,” in EMNLP. Association for Computational Linguistics, 2022, pp. 3781–3797.
127. S. Cho, S. Jeong, J. Seo, and J. C. Park, “Discrete prompt optimization via constrained generation for zero-shot re-ranker,” in ACL (Findings). Association for Computational Linguistics, 2023, pp. 960–971.
128. W. Sun, L. Yan, X. Ma, P. Ren, D. Yin, and Z. Ren, “Is chatgpt good at search? investigating large language models as re-ranking agent,” CoRR, vol. abs/2304.09542, 2023.
129. X. Ma, X. Zhang, R. Pradeep, and J. Lin, “Zero-shot listwise document reranking with a large language model,” CoRR, vol. abs/2305.02156, 2023.
130. Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang et al., “Large language models are effective text rankers with pairwise ranking prompting,” arXiv preprint arXiv:2306.17563, 2023.
131. F. Ferraretto, T. Laitz, R. de Alencar Lotufo, and R. F. Nogueira, “Exaranker: Explanation-augmented neural ranker,” CoRR, vol. abs/2301.10521, 2023.
132. L. Boytsov, P. Patel, V. Sourabh, R. Nisar, S. Kundu, R. Ramanathan, and E. Nyberg, “Inpars-light: Costeffective unsupervised training of efﬁcient rankers,” CoRR, vol. abs/2301.02998, 2023.
133. A. Askari, M. Aliannejadi, E. Kanoulas, and S. Verberne, “Generating synthetic documents for crossencoder re-rankers: A comparative study of chatgpt and human experts,” CoRR, vol. abs/2305.02320, 2023.
134. H. Wachsmuth, S. Syed, and B. Stein, “Retrieval of the best counterargument without prior topic knowledge,” in ACL (1). Association for Computational Linguistics, 2018, pp. 241–251.
135. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “REALM: retrieval-augmented language model pretraining,” CoRR, vol. abs/2002.08909, 2020.
136. P. S. H. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler¨ , M. Lewis, W. Yih, T. Rocktaschel,¨ S. Riedel, and D. Kiela, “Retrievalaugmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.
137. W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W. Yih, “REPLUG: retrieval-augmented black-box language models,” CoRR, vol. abs/2301.12652, 2023.
138. G. Izacard, P. S. H. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, “Few-shot learning with retrieval augmented language models,” CoRR, vol. abs/2208.03299, 2022.
139. A. Lazaridou, E. Gribovskaya, W. Stokowiec, and N. Grigorev, “Internet-augmented language models through few-shot prompting for open-domain question answering,” CoRR, vol. abs/2203.05115, 2022.
140. H. He, H. Zhang, and D. Roth, “Rethinking with retrieval: Faithful large language model inference,” CoRR, vol. abs/2301.00303, 2023.
141. J. Liu, J. Jin, Z. Wang, J. Cheng, Z. Dou, and J. Wen, “RETA-LLM: A retrieval-augmented large language model toolkit,” CoRR, vol. abs/2306.05212, 2023.
142. O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham, “Incontext retrieval-augmented language models,” CoRR, vol. abs/2302.00083, 2023.
143. H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, “Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 10 014–10 037.
144. Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. DwivediYu, Y. Yang, J. Callan, and G. Neubig, “Active retrieval augmented generation,” CoRR, vol. abs/2305.06983, 2023.
145. S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatﬁeld-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. E. Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan, “Language models (mostly) know what they know,” CoRR, vol. abs/2207.05221, 2022.
146. Z. Jiang, J. Araki, H. Ding, and G. Neubig, “How can we know When language models know? on the calibration of language models for question answering,” Trans. Assoc. Comput. Linguistics, vol. 9, pp. 962–977, 2021.
147. O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis, “Measuring and narrowing the compositionality gap in language models,” CoRR, vol. abs/2210.03350, 2022.
148. O. Khattab, K. Santhanam, X. L. Li, D. Hall, P. Liang, C. Potts, and M. Zaharia, “Demonstratesearch-predict: Composing retrieval and language models for knowledge-intensive NLP,” CoRR, vol. abs/2212.14024, 2022.
149. O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant, “Answering questions by metareasoning over multiple chains of thought,” CoRR, vol. abs/2304.13007, 2023.
150. Y. Qin, Z. Cai, D. Jin, L. Yan, S. Liang, K. Zhu, Y. Lin, X. Han, N. Ding, H. Wang, R. Xie, F. Qi, Z. Liu, M. Sun, and J. Zhou, “Webcpm: Interactive web search for chinese long-form question answering,” in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 8968–8988.
151. S. MacAvaney, C. Macdonald, R. Murray-Smith, and I. Ounis, “Intent5: Search result diversiﬁcation using causal language models,” CoRR, vol. abs/2108.04026, 2021.
152. N. Craswell, “Mean reciprocal rank,” in Encyclopedia of Database Systems, L. Liu and M. T. Ozsu,¨ Eds. Springer US, 2009, p. 1703.
153. K. Jarvelin and J. Kekal ainen, “Cumulated gain-based evaluation of IR techniques,” ACM Trans. Inf. Syst., vol. 20, no. 4, pp. 422–446, 2002.
154. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 2002, pp. 311–318.