大语言模型时代的信息检索综述

庞亮，邓竞成，顾佳，沈华伟，程学旗
中国科学院计算技术研究所，智能算法安全重点实验室，北京，100190
中国科学院大学，北京，100190
pangliang@ict.ac.cn, dengjingcheng23s@ict.ac.cn

摘要

以大语言模型为代表的生成式人工智能迅猛发展，标志着人工智能从判别时代向生成时代的转变。这一进步极大地推动了信息检索技术的发展，本文对大语言模型对信息检索领域的影响进行了深入的综述。从性能改进到模式颠覆，逐步展开论述大语言模型对信息检索领域的影响。针对传统信息检索流程，大语言模型凭借强大的语义理解和建模能力，显著增强索引、检索和排序等信息检索模块的性能。同时，文章也探讨了大语言模型可能取代传统信息检索的趋势，并催生了新的信息获取方式，或将是新一次信息时代的寒武纪。此外，大语言模型对内容生态的深远影响也值得关注。

关键词： 信息检索；大语言模型；索引；检索；排序

A Review of Information Retrieval in the Era of Large Language Models

PANG Liang, DENG Jingcheng, GU Jia, SHEN Huawei, CHENG Xueqi
CAS Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences / Beijing, 100190
University of Chinese Academy of Sciences / Beijing, 100190
pangliang@ict.ac.cn, dengjingcheng23s@ict.ac.cn

Abstract

The rapid development of generative artificial intelligence, exemplified by large language models, signifies a shift in artificial intelligence from the era of discrimination to generation. This advancement has greatly propelled the development of information retrieval technology. This paper provides an in-depth review of the impact of large language models on information retrieval. It discusses the influence of large language models on information retrieval, from performance improvements to paradigm shifts. In the context of traditional information retrieval processes, large language models significantly enhance the performance of various information retrieval modules such as indexing, retrieval, and ranking, owing to their powerful semantic understanding and modeling capabilities. Additionally, the paper explores the potential trend of large language models replacing traditional information retrieval methods and giving rise to new ways of obtaining information, possibly marking a new Cambrian explosion in the information age. The profound impact of large language models on the content ecosystem is also worthy of attention.

Keywords: Information Retrieval , Large Language Models , Indexing , Retrieval , Ranking

第二十三届中国计算语言学大会论文集，第98页-第119页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

信息检索是指从海量的文本、数据或多媒体中，根据用户需求找出相关信息，并呈现给用户的过程。在当今信息爆炸的时代，信息检索的价值愈发显著，是人们获取信息的重要途径(Robertson et al., 2009)。传统的信息检索通常包含索引、检索、排序等模块。索引模块（Indexing）为海量数据构建快速的访存机制，利用更少的空间，存储更多的数据，构建更快的访问(Negi et al., 2012)；检索模块（Retrieval）为用户需求筛选候选文档，保证相关信息的召回率(Xu et al., 2023d)；排序模块（Ranking）为用户需求提供精确的排序策略，确保更相关的文档排在更靠前的位置(Pang et al., 2017)，最后以列表的形式展现给用户，用户通过从上到下浏览文档列表，满足用户的信息需求。信息检索规范化的流程定义，让其快速规模化，诞生了类似谷歌、百度、必应等企业，但也面临着非常棘手的挑战，包括复杂用户需求的语义理解、全系统流程的相关性优化、单调不直接的列表信息形式等，有待解决。

近年来大语言模型（LLMs）的迅速发展，以ChatGPT为代表的模型(Brown et al., 2020)，基于Transformer架构(Vaswani et al., 2017)，在开放域问答(Xu et al., 2024a)、数学解题(Zhao et al., 2024)、对话系统(Zhou et al., 2023)、机器翻译(Piergentili et al., 2024)、文本摘要(Pakull et al., 2024)等领域表现出色。大语言模型凭借其强大的语义理解能力、顺畅的多轮交互能力、丰富的生成展现形式，已然成为各领域的研究热点，而利用大语言模型的优势来优化信息检索的各个模块(Zhu et al., 2023)，例如索引模块、检索模块、排序模块，成为最直接的应用目标。基于大语言模型的细粒度嵌入向量可以提升索引阶段的信息区分度，深度用户查询理解和重构可以提升检索阶段的信息召回率，内部充分交互后的生成式文档重排可以提升排序阶段的精确度。

除了嵌入在传统信息检索系统流程中提升效果，近年来还涌现了大量以大语言模型主导的信息检索新范式(Ma et al., 2024)，颠覆传统的信息检索流程。新型搜索引擎大致可以分为两类，一类是大语言模型作为代理进行网页浏览检索并整理检索结果的代理式检索，另一类是利用大语言模型的对话生成能力，将大语言模型作为搜索引擎的核心部分的交互式检索。与传统的关键字搜索引擎不同，基于大语言模型的新型搜索引擎更注重理解用户的意图和上下文，并提供与之相关的搜索结果。这类搜索引擎在用户与系统交互时提供有针对性的信息检索，能够为用户提供更智能、更个性化、更高效的信息检索体验，从而改善用户与系统之间的交互和沟通(Spatharioti et al., 2023)。广告为搜索引擎带来巨大的经济效益，而随着这一类新型搜索引擎的兴起，在传统搜索引擎中嵌入的广告应如何融入新型搜索引擎则是一个具有应用前景的话题(Feizi et al., 2023)。目前的主要做法是将广告作为文本嵌入大语言模型生成的内容。

大语言模型推动了信息检索领域的发展，也对信息检索的内容生态带来了深远影响，其中包括偏见问题、不公平问题和创作消费问题等。研究者们发现基于大语言模型的信息检索更倾向于检索人工智能生成的内容(Dai et al., 2023a; Xu et al., 2023c)，大语言模型生成的错误内容、不公平内容以及过多的人工智能创作对信息检索生态的影响(Dai et al., 2024)。

本文的结构如图 1所示，首先我们在第2节概述信息检索的主要模块以及大语言模型的发展总结。其次，我们将在第3节介绍大语言模型在传统信息检索中的应用，在第4节总结以大语言模型为核心的信息检索新范式，在第5节详述大语言模型生成的内容对信息检索生态的影响。最后，在第6节中讨论信息检索结合大语言模型在未来的发展方向，并在第7节对本文进行总结。

2 背景

2.1 信息检索

信息检索系统旨在从大规模的信息资源中根据用户的需求和查询提供相关的信息。它的主要作用是帮助用户快速准确地获取所需的信息，从而满足他们的信息需求。在本文中我们主要关注文本模态的信息检索系统，并将其分为三大块，分别为文本索引、文本检索和语义排序。

2.1.1 索引模块

文本索引是一种用于组织和加速文本数据检索的数据结构。它的主要功能是将文本数据中的关键词和它们出现的位置建立关联，以便在搜索过程中能够快速定位和检索相关文档。根据信息检索的两大范式——稀疏检索和稠密检索，我们可以将文本索引划分为两个主要部分。针对稀疏检索的文本索引被认定为是一种用于组织和加速文本数据检索的数据结构(Robertson et al., 1994)。它通常包括分词、去除停用词、正规化和词干提取等步骤，最终构建类似于倒

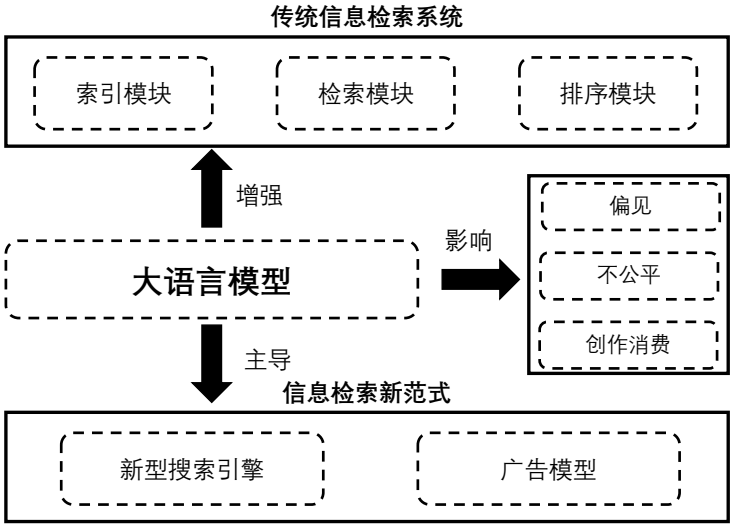

图 1: 大语言模型对现代信息检索系统的改变。主要分为增强传统信息检索系统、主导信息检索新范式和对检索生态的影响。

排索引的数据结构。在稠密检索中，人们一般通过文本嵌入模型来构建文本索引(Izacard and Grave, 2021)。此时文本索引被视为低维空间中的稠密向量。

2.1.2 检索模块

文本检索是指通过对文本内容的处理和分析，从大规模的文本集合中匹配和提取与用户查询相关的文档或文本片段。文本处理主要涉及查询处理和匹配这两种方法。前者对查询进行分析、解析和规范化。这包括去除停用词（如"和"、"的"等常见词汇），处理查询的语法和语义，以及提取查询中的关键词和条件。而后者通常是指使用诸如倒排索引（inverse index）等技术来快速定位包含查询关键词的文档。匹配完成后，搜索结果会按照相关性进行排序，以便将最相关的文档排在前面。此外随着深度学习技术的兴起，最近的研究(Izacard and Grave, 2021)主要围绕将查询和文档映射到各向同性的向量空间中，然后通过内积计算来计算它们的相关性分数。这种范式转变可以更有效地捕获查询和文档的语义相似性。总的来说这一模块充当信息检索系统中的首轮文档检索器，它从大规模文档集合中召回广泛相关文档作为候选。因此它定位相关文档的效率对于信息检索系统尤为重要。

2.1.3 排序模块

语义排序在一些研究中又被称之为重排序，它是信息检索中的另一个关键模块。与文本检索阶段强调效率和性能的平衡不同，它主要对召回的文档进行细粒度重新排序。为了提高排序质量，最近的研究(Xiao et al., 2023)提出了比传统内积匹配更复杂的方法，从而为模型提供更加丰富的语义匹配信号。

2.2 大语言模型

与传统语言模型不同，生成式大语言模型在处理复杂任务方面具有出色表现，这被称之为涌现能力(Wei et al., 2022)。最著名的大语言模型是来自OpenAI的ChatGPT(Ouyang et al., 2022)，它是大模型时代的一个里程碑。现有的大语言模型根据架构可以分为两类，第一类是编码器-解码器模型(Raffel et al., 2020; Zeng et al., 2023)，第二类是解码器模型(Ouyang et al., 2022)。编码器-解码器模型将文本输入编码器中转为向量，然后将向量输入解码器以获得输出。具有代表性的模型有T5(Raffel et al., 2020)和GLM(Zeng et al., 2023)。而更多的还是以GPT系列为代表的解码器模型，它们依赖于Transformer的解码器架构，从左至右自回归式地生成单词或字。具有代表性的模型有InstructGPT(Ouyang et al., 2022)、LLaMA系列模型(Touvron et al., 2023a; Touvron et al., 2023b)等。

3 信息检索系统中的大语言模型应用

本节主要描述在传统信息检索系统中大语言模型能够带来的改进。

3.1 大语言模型增强索引模块

针对稀疏检索的文本索引已经相当成熟，大语言模型对其带来的改进并不明显。而针对稠密检索的文本索引，大语言模型的出现为其性能带来了显著提升，因此我们重点关注这部分。在本文中，我们将大语言模型对文本索引的改进分为两个主要部分：生成训练数据和构造索引向量。

3.1.1 利用大语言模型生成训练数据

从模型训练角度看，大量高质量的训练数据至关重要，这可以使文本嵌入模型获得全面的语义知识和准确的语义空间。不幸的是，收集人工注释的相关性标签既耗时又昂贵。它限制了文本嵌入模型的知识边界及其跨不同应用领域进行泛化的能力。考虑到大语言模型强大的文本理解和生成能力，许多研究者利用基于大语言模型驱动的流程构建相关的查询-文档对以扩大文本嵌入模型的训练数据。

考虑到在现实世界中文档远比查询更加丰富，因此InPars(Bonifacio et al., 2022)提出使用类似于GPT-3的大语言模型来针对未被标注的文档生成相应伪查询。具体而言，它利用GPT-3强大的上下文学习能力（In-Context Learning）把一些查询-文档对作为示例输入模型。随后GPT-3针对给定文档生成可能的伪查询。因此利用这种方法可以轻松合成大量的训练数据。然而由于InPars方法较为简单，容易生成一些带有噪声或不相关信息的查询-文档对，因此InPars-v2(Jeronymo et al., 2023)在其之上使用更强大的大语言模型，并加入了现有的强大的重排序器对生成结果进行筛选，从而确保了最终数据的质量。除了使用重排序器过滤生成的样本对，也有研究专注于优化生成伪查询阶段的性能。AugTriever(Meng et al., 2022)提出两种构造查询-文档对的新策略，分别为查询提取和转移查询生成。查询提取针对文档中的某一跨度生成精确查询，而转移查询生成为其他自然语言处理任务（如文本摘要）也生成伪查询。这两种策略极大增强了生成的样本对的性能。UDAPDR(Saad-Falcon et al., 2023)采用了两阶段的伪查询生成方法，它首先采用强大的大语言模型生成少量高质量的伪查询，然后把这些高质量的查询-文档对作为普通的大语言模型的输入示例以生成大量伪查询。这种方法平衡了计算成本和生成质量。除了关注生成质量外，也有部分工作关注生成的任务类型和生成的数量。(Ma et al., 2023a)等人利用对比学习和瓶颈查询生成将大语言模型当中的知识有效地传递给文本嵌入模型，此外他们还结合了课程学习策略来减少对大语言模型推理的依赖。最终这些生成的数据被用于预训练文本嵌入模型，并且取得了出色的性能。Gecko(Lee et al., 2024b)扩大了合成的数据量并生成了更广泛的任务类型。然后它为每个查询检索一组候选段落，并使用相同的大语言模型重新标记正例段落和硬负例段落，进一步优化数据质量。这步操作保证它使用的大部分数据来源于真实数据而非合成数据，从而减轻合成数据带来的偏差。Promptagator(Dai et al., 2023b)专注于对特定检索任务合成数据，然后使用此数据训练文本嵌入模型，在11个不同的检索任务上平均nDCG上升1.2%。(Wang et al., 2024a)等人利用专有的大语言模型为近100种语言的数十万个文本嵌入任务生成各种合成数据，然后利用对比学习损失微调Mistral-7B模型，在BEIR(Thakur et al., 2021)和MTEB(Muennighoff et al., 2023)基准上取得了先进的性能。

3.1.2 利用大语言模型构造索引向量

传统的文本嵌入模型大部分是编码器结构的。虽然有少部分研究者尝试使用编码器-解码器结构或者仅解码器结构的模型产生嵌入向量，但是效果不佳。大语言模型的出现为使用仅解码器结构的模型产生高质量嵌入向量提供可能性。(Ma et al., 2023b)等人为每个文档拼接上一个终止符‘< \s >’，接着将它们输入LLaMA-2模型当中，并使用终止符‘< \s >’作为对应文档的嵌入表示。然后它采用InfoNCE损失函数对模型进行端到端优化，最终构造出RepLLaMA模型。实验证明大语言模型有成为稠密检索器的能力。(Li et al., 2023a)等人提出两个可以将大语言模型调整为稠密检索当中嵌入编码器的任务，分别叫做基于嵌入的自动编码（EBAE）和基于嵌入的自回归（EBAR）。它们确保大语言模型产生的文本嵌入可以重建输入语句的表示并预测下一语句的表示，大大提高了基于大语言模型的索引向量在各种稠密检索基准上的性能。考虑到自回归语言模型当中因果注意力机制的限制导致输入当中某一个标记表示向量并不能包含后续标记表示向量的信息，(Springer et al., 2024)等人提出一种叫做回声嵌入的方式，将文档重

复输入两次，并提取第二次出现的标记向量作为嵌入向量，在不修改模型结构的情况下缓解这个问题。而(BehnamGhader et al., 2024)等人提出一个简单的三阶段策略改变模型参数从而缓解这个问题，即1）启用双向注意力机制，2) 屏蔽下一个标记预测任务，3) 进行无监督对比学习。(Lee et al., 2024a)等人与其类似，他们在对比学习期间删除了原始大语言模型中的因果注意力机制，并加入一个潜在的注意力层来获取池化嵌入向量。他们提出的模型目前在MTEB排行榜上展示出最好的表现。

3.2 大语言模型增强检索模块

构造文本索引之后，信息检索到了文本检索阶段。此时给出查询可以召回与之最相似的一批文档。这一阶段主要注重于检索效率和高召回率，它们维持搜索引擎性能和最终结果生成至关重要。最近的工作可以被分为两类。第一类为查询重写，它们主要利用具有出色理解和生成能力的大语言模型重新表述原始查询以解决查询表述歧义、不清晰、不完整以及查询和文档之间词汇不匹配等问题。第二类工作被称为生成式检索，它们将大语言模型当作索引库，针对查询直接生成索引。

3.2.1 查询改写

最经典的工作为Query2doc(Wang et al., 2023a)，它利用大语言模型根据原始查询生成相关段落，然后将原始查询和相关段落进行拼接，一同去召回相关文档。由于生成的段落包含额外的详细信息，这可以缓解词汇不匹配问题，同时对不明确的、简短的查询尤为有效。(Jagerman et al., 2023)等人为了获得高质量的伪相关段落，他们分析了在零样本、少样本和思维链设置下大语言模型生成伪相关段落的质量，最后发现在思维链设置下用大语言模型进行查询改写的效果最好。除此之外，还有大量的工作是有不同的方法扩展原始查询中的知识。(Feng et al., 2023)等人提出了InteR，它允许现代检索模型和大语言模型之间的协同作用来促进信息细化。(Shen et al., 2023)等人建议通过使用查询和查询域内候选的组合来提示大语言模型，从而用其潜在答案来增强查询。(Lei et al., 2024)等人考虑到大语言模型的内在知识有限从而导致出现幻觉和信息过时等问题。因此他们提出引入语料库引导的查询扩展(CSQE) 来促进语料库中嵌入知识的整合。相关实验表明，CSQE无需任何训练即可表现出强大的性能，尤其是对于大语言模型缺乏知识的查询。针对法律领域的查询，(Tang et al., 2023b)等人使用大语言模型将复杂的和法律有关的查询简化为更易于搜索的法律事实和问题，然后采用基于提示的编码方案来进行有效的语言模型编码。然而，使用伪相关段落可能会给原始查询带来噪声并出现概念漂移，(Anand et al., 2023)等人利用重写的查询和文档对文本嵌入模型进行微调，让模型性能得到了极大的提升。

3.2.2 生成式检索

生成式检索将整个检索过程建模为生成式任务，具体可以分为四步，包括：（1）查询制定，即确定生成模型的输入；（2）文档标识符制定，即用短的标识符表示文档；（3）模型训练；（4）模型推理。在生成式检索中查询制定步骤一般比较简单，大部分工作均使用原始查询作为生成模型的输入。而模型推理步骤又和模型训练相关联。因此我们主要总结最近有关文档标识符制定和模型训练的研究。

理论上讲，生成式检索应该直接生成和查询对应的文档，然而由于生成模型上下文长度的限制，现有方法通常依赖于使用标识符来表示文档。最常见的是为语料库中的每一个文档分配一个数字标识符(Tay et al., 2022; Li et al., 2024a; Nadeem et al., 2022; Wang et al., 2022)。然而由于数字标识符缺乏语义含义，使其泛化性差。并且每当语料库更新时，新增文档标识符的构建以及记忆难度均会增加。使用类似于文档标题作为标识符是另一种解决方案(Chen et al., 2022; Cao et al., 2021; Lee et al., 2022a; Li et al., 2023b)，它将文档语义信息也纳入了标识符当中，在语义上可以和文档建立一对一的对应关系。然而这类方法因为很难设定段落标识符从而在更细粒度的段落级检索中表现不佳。并且在面向网页检索的任务当中也难以为网页构建高质量的标题标识符。除了这些之外，还有许多其他的建模方案，比如将文档中与查询语义相关的N元语法视为潜在的标识符(Bevilacqua et al., 2022; Chen et al., 2023a; Wang et al., 2023c),或者构建一个所谓的密码本，在根据文档内容学习最佳标识符(Sun et al., 2023a; Yang et al., 2023)。

生成式检索模型的训练一般分为两个阶段，分别为查询到标识符的训练和文档到标识符的

训练。对于缺乏语义表示的数字标识符和密码本标识符而言，后者的训练过程尤为重要。以上两个训练阶段均为生成式任务，然而也有工作(Li et al., 2024b)表明判别式训练的重要性。许多工作(Tang et al., 2024b; Zeng et al., 2024)证明了在生成式检索模型上引入判别式训练的有效性。

得益于大语言模型出色的理解能力和生成能力，还有不少研究脱离了传统生成式检索范式。比如(Ziems et al., 2023a)等人将大语言模型视为内置搜索引擎。他们直接构建用于文档检索的URL指令，并发现当提供一些例子给大语言模型时，大语言模型可以生成Web URL，其中近90%的相应文档包含对开放域问题的正确答案。(Yu et al., 2023)等人用大语言模型替换传统的文档检索器，直接根据给定查询生成上下文文档，在TriviaQA(Joshi et al., 2017)和WebQ(Berant et al., 2013)两个数据集上上分别获得了71.6和54.4的精确匹配分数。

3.3 大语言模型增强排序模块

语义排序作为信息检索系统中的最后一个阶段，旨在根据语义相关性对召回的文档重新排序。我们将在此阶段中使用大语言模型的工作分为两类，分别为利用大语言模型监督微调重排序器和利用大语言模型进行生成式排序。

3.3.1 监督微调重排序器

与利用大语言模型生成数据来监督微调检索器类似，监督微调重排序器也是一个很常见的做法。比如(Ferraretto et al., 2023)等人使用诸如GPT-3.5之类的大语言模型，通过解释来增强检索数据集，并训练一个序列到序列的排序模型，以输出给定查询-文档对的相关性标签和解释。实验证明在这种方法生成几千个样本上进行微调，性能与没有解释的使用3倍甚至更多样本进行微调的模型相当。(Boytsov et al., 2023)等人借鉴Inpars(Bonifacio et al., 2022)的思路，构造出一个轻量级的被称为InPars-Light的方法。它提示大语言模型生成合成查询-文档对，并针对比之前小7到100倍参数的排序模型进行无监督训练，最终在五个英文检索数据集上取得显著性改进。与注重于生成合成查询的工作不同，也有一部分工作强调生成合成文档的重要性。(Boytsov et al., 2023)等人构造出一个名为ChatGPT-RetrievalQA的数据集，它基于大语言模型响应用户查询生成合成文档来构建。他们利用此数据集和人工生成的数据微调了一系列重排序器。在多个数据集上的结果证明使用此数据集训练的重排序器比用真实数据训练的重排序器在统计上更有效。(Askari et al., 2023)等人提出DocGen和DocGen-RL两种方法，前者从查询生成合成文档，后者利用强化学习进一步优化DocGen，从而提高生成的合成文档与其对应查询之间的相关性。

3.3.2 生成式排序

生成式排序将排序任务视为生成式任务进行建模。在大语言模型出现之前，这个领域已有许多工作(Nogueira et al., 2020; Ju et al., 2021; Pradeep et al., 2021; Zhuang et al., 2023b)。而强大的大语言模型出现为这个领域又带来新的可能。按照生成方式我们将它们分为逐个生成（pointwise）、列表生成（listwise）、成对生成（pairwise）和集合生成（setwise）四大类。

逐个生成方法是指每次给定一个查询-文档对，生成式重排序模型生成它们的相关性得分。RankLLaMA(Ma et al., 2023b)将查询-文档对以“query: query document: document [EOS]”的模板输入到LLaMA模型中，并取“[EOS]”标记的最后一层嵌入表示进行相关性得分计算。由于此任务和大语言模型的预训练任务形式差距过大，因此在推理前需要对模型进行微调。(Zhuang et al., 2023a)等人认为生成二进制相关性标签（如‘True’或‘False’）的方法由于缺少中间相关性标签选项可能会导致大语言模型为与查询部分相关的文档提供嘈杂或有偏见的答案，因此他们将细粒度相关性标签合并到生成式重排序器的提示中，使它们能够更好地区分与查询具有不同相关性级别的文档，从而得出更准确的排名。与让重排序器生成相关性标签的工作相反，也有一部分工作仅给定文档，然后计算基于此文档生成实际查询的平均对数似然分数来确定此查询-文档对的相关性分数。(Sachan et al., 2022)等人直接计算以文档为条件的输入查询的概率，在完全开放域问答上取得了最先进的性能。(Zhuang et al., 2023c)等人重点研究了近期大语言模型真正的零样本查询似然排名有效性。同时他们还引入了一种新颖的先进排名系统，该系统将基于大语言模型的查询似然模型与混合零样本检索器相结合，在零样本和小样本场景中都表现出卓越的有效性。

列表生成方法将查询和所有相关文档输入生成式重排序器，然后直接让其按照相关性顺序

输出文档标识符。(Ma et al., 2023c)等人采用零样本列表生成方法在三个网络搜索数据集上进行实验，结果表明零样本列表生成方法不仅在对第一阶段检索结果进行重排序时优于零样本逐个生成方法，而且还可以充当最终阶段的重排序器，以改进逐个生成方法的排名结果，从而提高效率。(Pradeep et al., 2023)等人发布了第一个完全开源的生成式重排序大语言模型，叫做RankVicuna。它能够在零样本设置中执行高质量的列表式重排序，并取得与基于GPT-3.5的零样本列表重排序相当的有效性。然而列表生成方法对输入中的文档顺序异常敏感(Sun et al., 2023b)，当文档顺序随机打乱时它的效果甚至不如BM25。为此(Tang et al., 2023a)等人提出了一种被叫做排列自洽的方式，它的主要思想是边缘化提示中的不同列表顺序，以产生具有较少位置偏差的顺序无关的排名。他们从理论上证明了此方法的稳健性，表明在存在随机扰动的情况下可以收敛到真实排名。最终的效果超越了之前列表重新排序的最高水平。

成对生成方法每次给定一个查询和两个文档，然后要求生成式重排序模型生成相关性更高的文档标识符。最后再采用排序算法对所有相关文档进行重排序。(Qin et al., 2023b)等人认为现有的大语言模型无法理解逐一生成和逐列表生成的方式。因此他们使用成对排名提示(PRP)的新技术来显著减轻大语言模型的负担，并使用中等规模的开源大语言模型在标准基准上达到了最佳排名。

集合生成方法每次给定一个查询和一些文档，然后要求生成式重排序模型生成最相关的文档标识符或利用所有文档标识符的logits进行相关性排序。(Zhuang et al., 2023d)等人首次提出setwise提示方法，这是对之前三种方法的补充。通过在一致的实验框架内进行比较评估，并考虑模型大小、生成消耗、延迟等因素，他们表明setwise方法本质上在有效性和效率之间的平衡。比如pointwise方法在效率方面得分很高，但其有效性较差。相反，pairwise 方法表现出卓越的有效性，但会产生高计算开销。而setwise方法减少了排名过程中大语言模型的推理次数。这显著提高了基于大语言模型的零样本排名的效率，同时还保持了较高的零样本排名有效性。

4 大语言模型主导的信息检索新范式

本章节主要介绍以大语言模型为主导的信息检索新范式，主要包含新型搜索引擎，区别于基于传统信息检索方式的传统搜索引擎，以及在新型搜索引擎中广告模型应该如何适应。

4.1 新型搜索引擎

基于大语言模型的新型搜索引擎充分利用了大语言模型的生成能力，为用户提供了更好的检索体验。(Ziems et al., 2023b)发现大语言模型可以遵循人类的指令，直接生成用于文档检索的URL，大语言模型可以被认为是内置的搜索引擎。(Tang et al., 2024a)等人提出了一个端到端并且由大语言模型驱动的信息检索架构——自检索，其中IR系统所需的能力可以完全内化到一个单一的大语言模型中，并在IR过程中深入利用大语言模型的能力。本节根据大语言模型扮演的不同角色分为代理式检索和交互式检索。

4.1.1 代理式检索

近年来，以大语言模型为核心的代理式检索方法逐渐受到关注和应用，展现出了巨大的潜力和实际效果。由于大语言模型的知识是有限的，因此出现了通过检索等方式结合外部知识提高大语言模型的生成能力，并称之为检索增强生成RAG(Lewis et al., 2020a)。由此，有部分研究通过网络检索来提高大语言模型的能力，在这一类检索中，首先从数据索引中检索相关条目，然后代理以大语言模型为核心，模仿人类浏览网页的操作处理检索到的条目，以使用模型进行最终预测。

WebGPT(Nakano et al., 2021)是OpenAI提出的一个创新方法解决长篇问答问题，利用Bing创建了一个基于文本的Web 浏览环境。这个系统中，经过微调的GPT-3语言模型作为检索代理，在环境中执行检索任务。WebGPT使用模仿学习和强化学习等一般方法，以端到端的方式改进了信息的检索和合成过程。此外，生成的答案中包含了来自网页段落的参考文献，从而提高了答案的可靠性和可信度。然而，WebGPT也存在一定的局限性。针对这些问题，WebGLM(Liu et al., 2023)提出了改进方案。WebGLM增强了大语言模型的网络搜索和检索功能，同时确保在实际部署中的效率。通过引入人类偏好的机制，WebGLM提高了模型在实际应用中的表现，使其能够更高效、更准确地满足用户的需求。WebShop(Yao et al., 2022)展示了一个更为复杂和真实的应用场景。WebShop开发了一个模拟电子商务网站环境，拥有118万种真实产品和12,087条众包文本指令。在这个环境中，代理需要浏览多种类型的网页，并根据

指令执行不同的操作，如查找、定制和购买产品。这不仅测试了语言模型在复杂交互任务中的表现，也为未来的实际应用提供了宝贵的经验。在中文信息检索领域，WebCPM(Qin et al., 2023a)则提出了第一个中文长篇问答（LFQA）数据集。WebCPM的信息检索基于交互式网络搜索，能够实时与搜索引擎交互。通过对预训练语言模型进行微调，WebCPM模仿了人类的网络搜索行为，根据收集到的事实生成答案。这种方法不仅提高了中文信息检索的效率，还增强了系统对复杂问答任务的处理能力。

针对基于检索的模型，(Basu et al., 2022)提出了一种正式处理方法来描述它们的泛化能力。通过分析这些模型在不同场景下的表现，研究揭示了它们在面对不同类型问题时的适应能力和局限性。这些理论研究为进一步优化和改进检索模型提供了重要的参考和指导。尽管大语言模型驱动的Web代理在信息检索方面展现了巨大潜力，但也面临着安全威胁。WIPI(Wu et al., 2024)介绍了一种新型威胁，能够间接控制Web代理执行嵌入在公开网页中的恶意指令。即使在黑盒环境下，这种方法仍然能够实现超过90%的攻击成功率，揭示了当前Web代理的安全漏洞。这为未来更安全的大语言模型系统设计提供了重要见解和方向。

以大语言模型为核心的代理式检索方法结合了大语言模型和信息检索，展示了巨大的潜力。通过不断改进和优化这些模型，我们可以更高效、更准确地利用大语言模型从海量数据中提取有用信息，从而满足用户日益增长的需求。

4.1.2 交互式检索

交互式检索（Interactive Information Retrieval, IIR）是一种通过多轮对话和反馈的方式进行信息检索的方法，旨在提升检索的精确度和用户满意度。这个过程强调用户在检索中的主动参与和系统对用户反馈的即时响应。通过多轮交互，系统能够更好地理解用户需求，提供更相关的结果。

但是在交互式检索的过程中，不仅需要从单个查询中理解用户的意图，更需要结合上下文，要求解锁模型有更强的上下文理解能力。由于大语言模型的迅速发展，其强大的自然语言处理能力和上下文理解能力，能够显著提升交互式检索系统的智能化水平和用户体验，使检索过程更加高效、精准和个性化。

和传统会话检索的区别 与传统的关键词搜索相比，IIR的交互性支持逐步深入的查询，能够更好地满足用户的复杂需求，并提供了更自然的用户体验，给信息检索领域带来了新的机遇和挑战(Vtyurina et al., 2017; Radlinski and Craswell, 2017)。因此，在大语言模型出现之前，已经存在多轮交互的信息检索研究，如会话检索，在传统的信息检索流程中引入多轮交互信息，提高信息检索的性能。Radlinski和Craswell(Radlinski and Craswell, 2017)提出了会话搜索的理论框架。在对话中的每一个来回步骤中，系统向用户提供一些信息，用户做出响应。在会话检索系统中，用户可以像与人对话一样与系统交互，提出问题并根据系统提供的结果进行进一步的提问和调整，经过多轮会话得到信息(Gao et al., 2023)。

作为大语言模型的前身，预训练语言模型（PLMs）如Bert、GPT-2等在大语言模型之前就曾被多次用于多轮交互的信息检索，在传统信息检索的步骤中起作用(Dalton et al., 2020; Voskarides et al., 2020)。又有研究建立模型通过检索恰当的澄清问题与用户交互以明确用户的查询需求，(Aliannejadi et al., 2019)等人制定了在开放域会话系统中搜索信息的问题澄清的任务，模型通过多轮主动询问来逐步明确用户需求。他们提出了一个检索框架，包括三个组成部分：问题检索，问题选择，和文档检索。而在此之后，(Zamani et al., 2020)分析了从Bing 搜索日志中采样的查询重构数据来确定开放域搜索查询的澄清分类法，以进一步研究为开放域搜索任务生成澄清问题。

可见会话检索仍然基于传统的信息检索流程，通常包含三个组件构成：上下文查询理解，文档检索（包含建立索引）和文档排名(Gao et al., 2023)。与一般的关键字信息检索相比，会话检索由多轮对话组成，需要更加强大的上下文理解能力，但由于之前技术的局限性，这一部分的研究具有一定的挑战性，难以大规模普及。而大语言模型出色的自然语言理解和上下文理解能力，为交互式检索带来新的突破。与会话检索不同，基于大语言模型的交互式检索采用新的信息检索范式，充分利用大语言模型的自然语言生成能力，采用生成问题而不是检索问题(Aliannejadi et al., 2019)的方式与用户对话，打破了原有的信息检索的流程框架。同时，其生成能力也为满足用户的新的检索意图（如，创作）提供了发展前景。

新一代可应用的交互式检索 大语言模型的发展为传统搜索引擎带来巨大的改变。以往只能输入一次关键字，搜索引擎返回大量排序后的搜索结果，需要用户再次鉴别结果，如

果检索失败，用户需要重新搜索(Maoro et al., 2024)。Microsoft曾揭示传统搜索引擎中的问题——几乎一半的网络搜索都得不到准确的答复。最近，大语言模型已与网络搜索相结合，以实现一种新的大语言模型驱动的交互式检索模式。此外，大语言模型驱动的搜索引擎如New Bing、Perplexity AI等能够理解以自然语言表达的复杂查询，使用户能够像在对话中一样直观地提问，为用户提供更准确的搜索结果，甚至可以提供创作和编写的功能(Ma et al., 2024)。

New Bing，通常称为Bing Chat或者简单称为Bing，是微软Bing搜索引擎的一次重大升级，结合了先进的人工智能功能，以增强搜索体验。New Bing集成了OpenAI的GPT-4等大语言模型。Bing提供实时更新信息和强大的视觉搜索功能，能够与其他微软服务和产品集成，创造了一个无缝的微软生态系统体验。Perplexity AI是一个基于大语言模型的搜索引擎，旨在通过理解问题的上下文并以对话形式提供精确、相关的信息来增强搜索体验。Perplexity AI的核心是先进的人工智能和机器学习算法，通过用户互动不断学习和改进。和New Bing一样，Perplexity AI 力求提供最新信息，使其成为获取当前事件和动态话题的有用工具。Bard是Google开发的一款基于人工智能的对话式搜索引擎，旨在利用自然语言处理技术提供更自然和互动的搜索体验。Bard支持多轮对话，用户可以在一次搜索中提出后续问题，Bard能够记住上下文并继续提供相关的回答。通过了解用户的偏好和兴趣，Bard可以提供个性化的搜索结果，You.com是一个新兴的搜索引擎，旨在通过整合人工智能和机器学习技术，提供个性化和隐私友好的搜索体验。You.com通过了解用户的偏好和兴趣，提供个性化的搜索结果，使用户能够快速找到最相关的信息。

(Ma et al., 2024)剖析了大语言模型驱动的交互式搜索引擎（特别是Bing Chat）为其生成的回答选择信息源的机制。研究表明，Bing Chat更偏好具可读性和分析性的源内容，而且其对文本的独特倾向是可以被底层大语言模型预测的。同时也揭示了RAG API和Bing Chat之间的一致文本偏好。(Gong and Cosma, 2023)介绍了一种新颖的跨模态搜索引擎Boon，它结合了两个最先进的网络：GPT-3.5-turbo大模型和VSE网络VITR，使用户能够执行图像到文本和文本到图像的检索并且能够就其选择的一个或多个图像进行对话。然而，(Wazzan et al., 2024)比较了传统搜索引擎和基于大语言模型的搜索引擎（Microsoft Bing Chat）在图像地理定位搜索任务中的性能，表明使用传统搜索的参与者比使用大语言模型搜索的参与者表现更好。在(Spatharioti et al., 2023)等人的研究中，基于大语言模型的搜索引擎的参与者能够更快地完成任务，并且参与者具有更满意的体验，但是如果大语言模型提供的信息不可靠，用户仍然会过度依赖错误信息。

在其他任务如推荐任务中，基于大语言模型的交互式检索也为其提供了新的研究范式,提供了超越传统推荐技术的更自然、更无缝的用户体验。(Huang et al., 2023)结合推荐模型和大语言模型各自的优势来创建一个多功能且交互式的推荐系统，它通过集成大语言模型，使传统推荐系统成为具有自然语言界面的交互式系统。为了改善企业网站上的搜索体验，(Maoro et al., 2024)提出了一个领域自适应的问答框架，结合了语义搜索和GPT-3.5，当返回答案时，用户可以提出后续问题并进行特定主题的交流，改善了企业网站的整体用户体验。(Völker et al., 2024)介绍了一种新颖的检索增强生成系统，该系统利用基于聊天的大语言模型来简化和增强出版物管理流程，使用户能够通过直观的聊天界面与各种网页平台如SemanticScholar等无缝交互。

4.2 广告模型

对于商业性质的查询，搜索引擎会提供相关的在线广告(Dubey et al., 2024)。在传统的搜索引擎中，通过关键词匹配广告，大语言模型能够提供更加准确的广告匹配。除此之外，随着大语言模型驱动的搜索引擎的发展，广告技术也经历了显著变化。传统的广告形式和投放方式可能不适用于大语言模型生成式的检索结果。

在传统搜索引擎中，广告与查询关键词匹配至关重要。传统技术在准确刻画查询和关键词之间的语义相关性方面存在局限性。为此，(Wang et al., 2024b)提出了一种基于大语言模型的关键词生成方法（LKG），能够一步到位地从搜索查询中提取相关关键词。这种方法利用大语言模型的语义理解能力，显著提高了关键词匹配的准确性和广告的投放效果。

广告的嵌入需要考虑到广告商出价等因素，因此需要一个新颖的拍卖机制来整合来自不同广告商的输入。将在线广告模型和拍卖框架转移到大语言模型环境中，带来了新的机遇和挑战。(Feizi et al., 2023)提出了一个合理的框架，包括修改大语言模型的原始输出、广告商为修改后的输出竞标、大语言模型计算广告的相关信息以及广告竞争并选择最终输出。这

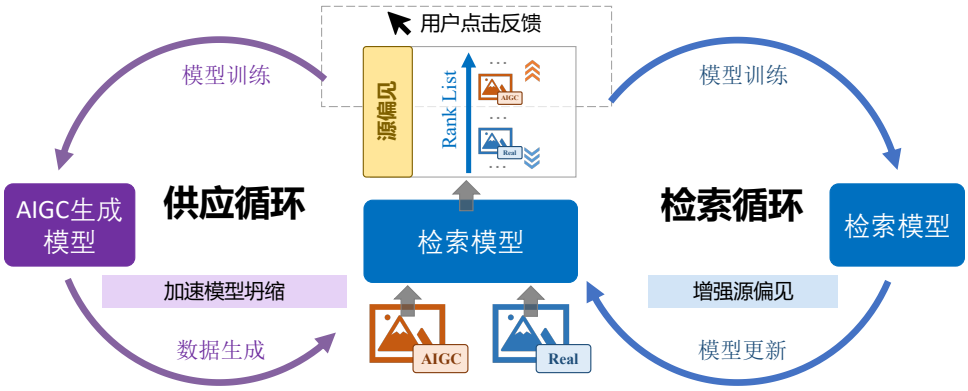

图 2: 源偏见问题示意图。IR模型倾向于将AI生成的图像排在真实图像之前，尽管它们具有非常相似的语义。这种偏见增加了生成的图像从互联网海量数据中暴露出来的可能性，这使得它们更容易被混入AIGC和检索模型的训练中，从而导致更严重的偏见并形成恶性循环。

种框架有效地整合了大语言模型的生成能力和广告竞价机制，提升了广告投放的精准度和效果。(Dütting et al., 2024)是第一篇引入大语言模型机制设计的论文，它提出了一种基于token的拍卖模型，该模型以大语言模型作为广告商代理，通过出价影响生成的广告内容。其中，广告段落是逐个token生成的，竞标者之间的出价在不同大语言模型中分布式聚合，从而产生优化的广告创意。与这样的方法不同，(Dubey et al., 2024)提出了一种分解框架，包括拍卖模块和大语言模型模块，其中分配和支付仍由拍卖模块决定，而大语言模型模块则根据拍卖模块分配的突出性生成广告摘要。这种方法不仅提高了广告的展示效果，还优化了广告内容的分配和生成。为了解决(Dütting et al., 2024)中广告商的支出会随着大语言模型生成的输出序列的长度而增长等限制，(Soumalias et al., 2024)引入了一种拍卖机制，不需要对大语言模型微调和访问模型的权重，就能汇总了多个广告商代理对用户查询回复的偏好需求，同时也为用户提供了有用的回答。

除了将广告融入大语言模型驱动的搜索引擎之外，(Schmidt et al., 2024)提到了对原生广告的检测。在大语言模型驱动的搜索引擎中，广告嵌入到生成的响应中，未来的用户很可能会面临生成的原生广告。研究表明，大语言模型也可以用于检测并阻止生成的原生广告，从而保护用户体验和信息的真实性。

大语言模型在搜索引擎广告中的应用，为广告技术带来了巨大的变化。通过引入智能拍卖机制、生成广告摘要等技术，广告商能够更精准地投放广告并适用于新的搜索引擎。

5 大语言模型对检索生态的影响

如第3节和第4节所述，大语言模型不仅对传统信息检索范式产生巨大影响，还产生了信息检索的新范式。这种变化带来新机遇的同时也带来了新的挑战，特别是大语言模型产生的偏见和不公平可能会威胁现在的检索生态。本节系统地研究了应用大语言模型可能会带来的偏见问题、公平性问题和创作消费问题。

5.1 偏见问题

大语言模型的出现让人们可以轻易地生成大规模合成数据，而这些合成数据会重溯检索数据的分布(Dai et al., 2024)。最近的研究(Dai et al., 2023a; Xu et al., 2023c)表明，现代检索模型，尤其是基于神经网络的模型，更倾向于检索由大语言模型生成的内容，而非人类创作的具有相似语义的内容。这种现象被称之为源偏见问题。如图 2所示。它的产生原因是，大语言模型生成的文本具有某种独特的嵌入表示。而神经检索模型可以捕捉这种表示，从而产生更高的排名。如果检索模型使用了合成数据进行训练，那么这种偏差会进一步放大。另外也有研究(Tan et al., 2024)表明，源偏见问题会从检索模型进一步延伸到生成模型。为了解决源偏见问题，目前的研究(Dai et al., 2023a; Xu et al., 2023c)主要集中于在检索模型的训练过程中引入去偏差约束。它们的思想在于从分布对齐角度将检索模型的相关性分布重新校正为理想状态，从而对不同来源的文档进行公平对待。

除了源偏见外，事实偏见也是另外一种常见的偏见问题，它被定义为大语言模型可能会产生与现实世界公认的事实信息不一致的内容(Dai et al., 2024)。(Lin et al., 2022; Lee et al., 2022c; McKenna et al., 2023)等人证明大语言模型会生成许多错误的答案，并且有可能欺骗人类。即使参数量更大的模型也无法避免这个问题。此外大语言模型在一些大规模基准(Chen et al., 2023b; Lee et al., 2022b; Deng et al., 2024; Wei et al., 2024a; Wei et al., 2024b)上的表现也验证了这种现象。事实偏见引入了大量非事实或"幻觉"内容。这种引入改变了检索数据的分布，从而导致检索过程中的偏差。目前有许多工作尝试缓解事实偏差。比如一部分(Gunasekar et al., 2023; Touvron et al., 2023b)侧重于为大语言模型提供高质量且事实正确的数据。还有一部分为在可信数据源中检索信息来增强大语言模型的生成(Ram et al., 2023; Lewis et al., 2020b; Shi et al., 2023; Deng et al., 2023; Xu et al., 2024d; Ding et al., 2024; Xu et al., 2024c)或者利用大语言模型自身的推理能力避免生成非事实问题(Xu et al., 2024b; Wang et al., 2023b; Chuang et al., 2023)。

5.2 不公平问题

(Dai et al., 2024)等人认为在信息检索系统中存在用户公平和项目（文档）公平两个概念，它们分别和社会学领域的平等和分配正义(Xu et al., 2023a)概念有关。具体来说用户公平被定义为信息检索系统应向不同用户提供公平和非歧视性的信息服务。而项目公平被定义为信息检索系统应该为较弱的项目提供更多的被检索到的机会。而大语言模型对检索生态会产生用户不公平和项目不公平的问题。

产生用户不公平的一个主要原因是检索数据中存在的歧视或者攻击性内容对特定群体产生了不成比例的影响。这些内容之所以会出现在检索数据中既可能是历史或者文化原因(Beukeboom and Burgers, 2019; Ntoutsi et al., 2020; Zhuo et al., 2023),也可能是大语言模型生成的(Fang et al., 2023)。此外，大语言模型之所以会生成这些内容，也源于它们的训练数据当中歧视性内容的文本(Beukeboom and Burgers, 2019)。以往的工作主要采用各种方法来过滤这些文本。比如(Ghanbarzadeh et al., 2023)等人通过性别调整来构造更公平的数据集来消除模型的偏见。(Xu et al., 2023b)等人经过大量的分析证实了大语言模型存在隐性的用户歧视，并强调识别和减轻隐性用户歧视的必要性。还有一些方法(Deldjoo and Noia, 2024; Ngo et al., 2021)使用降低包含歧视信息样本的重采样策略或者直接过滤和删除这些内容。此外，指令微调或基于人类反馈的强化学习也被证明可以让大模型对齐人类的价值观以有效促进公平性(Touvron et al., 2023b)。

产生项目不公平性问题的一个主要原因为某些项目的代表性不平衡导致在信息检索或评估过程当中的差异(Jiang et al., 2024)。此外大语言模型也有可能生成新的项目或文档，从而潜在引入了新的内容和观点(Das et al., 2024; Jr. and Licato, 2023)。为了减轻数据当中的项目不公平性，(Jiang et al., 2024)等人提出为不同的项目或文档进行重新加权以平衡项目或文档的代表性。

5.3 创作消费问题

大语言模型极大地降低了人们的创作门槛，让普通用户通过使用强大的内容生成模型（例如ChatGPT、Sora和GPT-4o）也能够创作高质量的文本或视频，这给在线内容生态系统和检索生态系统注入了新活力(Epstein et al., 2023)。然而，这种使用人工智能来进行创作消费的转型也带来了新的问题，即会导致市场过度饱和，个人创作者的内容更加难以被挖掘。另外，大语言模型并不是万能的，如第5.1和5.2节所述，它具有偏见和不公平的现象。如果高质量的人类创作内容被边缘化，那么依赖于广泛且多样化的数据集进行训练的大语言模型生成的质量必然下降(Yao et al., 2024)。因此探索人类生成内容和人工智能生成内容是否能稳定地以共生的形式存在是一个具有挑战性的方向。(Yao et al., 2024)等人拓展了Tullock竞争模型，从理论和实验上给出了一个具有希望的前景，即尽管生成式人工智能会扰乱人类内容生成的市场，但具有理想特征的稳定平衡是可以实现的。

6 未来发展

降低在信息检索系统中应用大语言模型的成本 大语言模型不仅可以改进传统的信息检索系统，还主导了信息检索的新范式。然而使用大语言模型会产生高昂的计算成本，尤其是对于

高校实验室或者是小规模的公司而言。即使是一些具有充足计算资源的大型互联网公司，当面临大量用户请求时也会产生巨大的成本压力。常见的解决方案包括大语言模型压缩和推理加速，但是这些方法有可能会损害大语言模型性能，从而对信息检索系统造成影响。因此需要开发更高效的大语言模型使用方式以应对成本挑战。

消除大语言模型生成内容的偏见 由于大语言模型生成的内容目前几乎被信息检索的所有阶段使用，然而这会产生偏见问题。它主要表现为改变了检索数据的分布，从而让检索模型更加倾向于检索大语言模型生成的内容。虽然目前已有一些工作使用去偏差约束来缓解此问题，但是它们无法彻底解决带有偏见的内容。

让大语言模型生成的内容更可信 由于大语言模型会产生幻觉，因此人们无法完全相信大语言模型根据用户查询生成的内容。有时这些内容看起来合理但实际上确实不合逻辑的甚至是虚假的。这为现代信息检索系统产生不利影响。因此需要正确认识到大语言模型在某些方面的局限性，从而开发出可信的现代信息检索系统。

7 结论

本文对大语言模型对信息检索领域的影响进行了深入的阐述。针对传统信息检索系统，大语言模型凭借出色的理解能力改变了索引模块、检索模块和排序模块。此外本文还探讨了大语言模型可能取代传统信息检索方法的趋势，并催生出新的信息检索范式，预示着信息时代的新发展。同时，本文也关注了大语言模型对内容生态的影响，包括偏见、不公平问题以及创作消费问题。此外，虽然将大语言模型应用到信息检索系统的前景广阔，但同时也带来了一系列挑战。未来的发展方向应从降低成本、消除偏见，提高可信度这几个方面去考虑。

参考文献

Mohammad Aliannejadi, Hamed Zamani, Fabio A. Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Abhijit Anand, Venktesh V, Vinay Setty, and Avishek Anand. 2023. Context aware query rewriting for text rankers using LLM. *CoRR*, abs/2308.16753.

Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, and Suzan Verberne. 2023. Expand, highlight, generate: Rl-driven document generation for passage reranking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10087–10099. Association for Computational Linguistics.

Soumya Basu, Ankit Singh Rawat, and Manzil Zaheer. 2022. Generalization properties of retrieval-based models. *CoRR*, abs/2210.02617.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *CoRR*, abs/2404.05961.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick S. H. Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *CoRR*, abs/2202.05144.

Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2023. Inpars-light: Cost-effective unsupervised training of efficient rankers. *CoRR*, abs/2301.02998.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: generative evidence retrieval for fact verification. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2184–2189. ACM.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2023a. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23. ACM, July.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023b. Complex claim verification with evidence retrieved in the wild. *CoRR*, abs/2305.11859.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023a. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *CoRR*, abs/2310.20501.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023b. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *CoRR*, abs/2404.11457.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC cast 2019: The conversational assistance track overview. *CoRR*, abs/2003.13624.

Debarati Das, Karin de Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, Ryan Koo, Jong Inn Park, Aahan Tyagi, Libby Ferland, Sanjali Roy, Vincent Liu, and Dongyeop Kang. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *CoRR*, abs/2401.14698.

Yashar Deldjoo and Tommaso Di Noia. 2024. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *CoRR*, abs/2403.05668.

Jingcheng Deng, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Regavae: A retrieval-augmented gaussian mixture variational auto-encoder for language modeling. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2500–2510. Association for Computational Linguistics.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Unke: Unstructured knowledge editing in large language models. *arXiv preprint arXiv:2405.15349*.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *CoRR*, abs/2402.10612.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. 2024. Auctions with LLM summaries. *CoRR*, abs/2404.08126.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. 2024. Mechanism design for large language models. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 144–155. ACM.

Ziv Epstein, Aaron Hertzmann, Laura Mariah Herman, Robert Mahari, Morgan R. Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, Jessica Fjeld, Hany Farid, Neil Leach, Alex Pentland, and Olga Russakovsky. 2023. Art and the science of generative AI: A deeper dive. *CoRR*, abs/2306.04141.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: An examination of news produced by large language models. *CoRR*, abs/2309.09825.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. 2023. Online advertisements with llms: Opportunities and challenges. *CoRR*, abs/2311.07601.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2023. Synergistic interplay between search and large language models for information re. *arXiv preprint arXiv:2305.07402*.

Fernando Ferraretto, Thiago Laitz, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2023. Exaranker: Explanation-augmented neural ranker. *CoRR*, abs/2301.10521.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural Approaches to Conversational Information Retrieval*, volume 44 of *The Information Retrieval Series*. Springer.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5448–5458. Association for Computational Linguistics.

Yan Gong and Georgina Cosma. 2023. Boon: A neural search engine for cross-modal information retrieval. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*, MM '23. ACM, October.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender AI agent: Integrating large language models for interactive recommendations. *CoRR*, abs/2308.16505.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Vitor Jeronymo, Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, Roberto de Alencar Lotufo, Jakub Zavrel, and Rodrigo Frassetto Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *CoRR*, abs/2301.01820.

Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. 2024. Item-side fairness of large language model-based recommendation system. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4717–4726. ACM.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Antonio Laverghetta Jr. and John Licato. 2023. Generating better items for cognitive assessments using large language models. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 414–428. Association for Computational Linguistics.

Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multi-view learning for passage re-ranking. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1803–1807. ACM.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022a. Generative multi-hop retrieval. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1417–1436. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022c. Factuality enhanced language models for open-ended text generation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *CoRR*, abs/2405.17428.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernández Ábrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024b. Gecko: Versatile text embeddings distilled from large language models. *CoRR*, abs/2403.20327.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 393–401. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle,

Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models A better foundation for dense retrieval. *CoRR*, abs/2312.15503.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023b. Generative retrieval for conversational question answering. *Inf. Process. Manag.*, 60(5):103475.

Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. 2024a. Towards a unified language model for knowledge-intensive tasks utilizing external corpus. *CoRR*, abs/2402.01176.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024b. Learning to rank in generative retrieval. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8716–8723. AAAI Press.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye, editors, *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4549–4560. ACM.

Guangyuan Ma, Xing Wu, Peng Wang, Zijia Lin, and Songlin Hu. 2023a. Pre-training with large language model-based document expansion for dense passage retrieval. *CoRR*, abs/2308.08285.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023b. Fine-tuning llama for multi-stage text retrieval. *CoRR*, abs/2310.08319.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023c. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting knowledge: Exploring the creative mechanisms of chat-based search engines. *CoRR*, abs/2402.19421.

Falk Maoro, Benjamin Vehmeyer, and Michaela Geierhos. 2024. Leveraging semantic search and llms for domain-adaptive information retrieval. In Audrius Lopata, Daina Gudonienė, and Rita Butkienė, editors, *Information and Software Technologies*, pages 148–159, Cham. Springer Nature Switzerland.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2758–2774. Association for Computational Linguistics.

Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Bhat, and Yingbo Zhou. 2022. Augtriever: Unsupervised dense retrieval by scalable data augmentation. *arXiv preprint arXiv:2212.08841*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Usama Nadeem, Noah Ziems, and Shaoen Wu. 2022. Codedsi: Differentiable code search. *CoRR*, abs/2210.00328.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Ambesh Negi, Mayur Bhirud, Suresh Jain, and Amit Mittal. 2012. Index based information retrieval system. *International Journal of Modern Engineering Research (IJMER)*, 2:945.

Helen Ngo, Cooper Raterink, João G. M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *CoRR*, abs/2108.07790.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernández, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tabea Margareta Grace Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A. Horn, and Christoph M. Friedrich. 2024. Wispermed at biolaysumm: Adapting autoregressive large language models for lay summarization of scientific articles. *CoRR*, abs/2405.11950.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17. ACM, November.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. *CoRR*, abs/2405.08477.

Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023a. Webcpm: Interactive web search for chinese long-form question answering. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8968–8988. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023b. Large language models are effective text rankers with pairwise ranking prompting. *CoRR*, abs/2306.17563.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *CoRR*, abs/2302.00083.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md. Arafat Sultan, and Christopher Potts. 2023. UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11265–11279. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Detecting generated native ads in conversational search. In Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, editors, *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 722–725. ACM.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large language models are strong zero-shot retriever. *CoRR*, abs/2304.14233.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Ermis Soumalias, Michael J. Curry, and Sven Seuken. 2024. Truthful aggregation of llms with an application to online advertising. *CoRR*, abs/2405.05905.

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *CoRR*, abs/2307.03744.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *CoRR*, abs/2402.15449.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023a. Learning to tokenize for generative retrieval. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *CoRR*, abs/2401.11911.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023a. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *CoRR*, abs/2310.07712.

Yanran Tang, Ruihong Qiu, and Xue Li. 2023b. Prompt-based effective input reformulation for legal case retrieval. In Zhifeng Bao, Renata Borovica-Gajic, Ruihong Qiu, Farhana Murtaza Choudhury, and Zhengyi Yang, editors, *Databases Theory and Applications - 34th Australasian Database Conference, ADC 2023, Melbourne, VIC, Australia, November 1-3, 2023, Proceedings*, volume 14386 of *Lecture Notes in Computer Science*, pages 87–100. Springer.

Qiaoyu Tang, Jiawei Chen, Bowen Yu, Yaojie Lu, Cheng Fu, Haiyang Yu, Hongyu Lin, Fei Huang, Ben He, Xianpei Han, Le Sun, and Yongbin Li. 2024a. Self-retrieval: Building an information retrieval system with one large language model. *CoRR*, abs/2403.00801.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. 2024b. Listwise generative retrieval models via a sequential learning process. *CoRR*, abs/2403.12499.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM, July.

Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 2187–2193, New York, NY, USA. Association for Computing Machinery.

Tom Völker, Jan Pfister, Tobias Koopmann, and Andreas Hotho. 2024. From chat to publication management: Organizing your related work using bibsonomy & llms. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '24. ACM, March.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022. A neural corpus indexer for document retrieval. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023c. NOVO: learnable and interpretable document identifiers for model-based IR. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 2656–2665. ACM.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *CoRR*, abs/2401.00368.

Yang Wang, Zheyi Sha, Kunhai Lin, Chaobing Feng, Kunhong Zhu, Lipeng Wang, Xuewu Jiao, Fei Huang, Chao Ye, Dengwu He, Zhi Guo, Shuanglong Li, and Lin Liu. 2024b. One-step reach: Llm-based keyword generation for sponsored search advertising. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1604–1608, New York, NY, USA. Association for Computing Machinery.

Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing traditional and llm-based search for image geolocation. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '24. ACM, March.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024a. Mlake: Multilingual knowledge editing benchmark for large language models. *CoRR*, abs/2404.04990.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. 2024b. Stable knowledge editing in large language models. *CoRR*, abs/2402.13048.

Fangzhou Wu, Shutong Wu, Yulong Cao, and Chaowei Xiao. 2024. WIPI: A new web threat for llm-driven web agents. *CoRR*, abs/2402.16965.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. 2023a. P-MMF: provider max-min fairness re-ranking in recommender system. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3701–3711. ACM.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2023b. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *CoRR*, abs/2311.07054.

Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2023c. Ai-generated images introduce invisible relevance bias to text-image retrieval. *CoRR*, abs/2311.14084.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023d. BERM: training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6620–6635. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1362–1373. ACM.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024b. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1362–1373. ACM.

Shicheng Xu, Liang Pang, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024c. List-aware reranking-truncation joint model for search and retrieval-augmented generation. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1330–1340. ACM.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024d. Unsupervised information refinement training of large language models for retrieval-augmented generation. *CoRR*, abs/2402.18150.

Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Auto search indexer for end-to-end document retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6955–6970. Association for Computational Linguistics.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. 2024. Human vs. generative AI in content creation competition: Symbiosis or conflict? *CoRR*, abs/2402.15467.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. 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.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. *Proceedings of The Web Conference 2020*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. 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.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1441–1452. ACM.

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. 2024. Exploring the compositional deficiency of large language models in mathematical reasoning. *CoRR*, abs/2405.06680.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Think before you speak: Cultivating communication skills of large language models via inner monologue. *CoRR*, abs/2311.07445.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. *CoRR*, abs/2310.14122.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023b. Rankt5: Fine-tuning T5 for text ranking with ranking losses. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2308–2313. ACM.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023c. Open-source large language models are strong zero-shot query likelihood models for document ranking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8807–8817. Association for Computational Linguistics.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2023d. A setwise approach for effective and highly efficient zero-shot ranking with large language models. *CoRR*, abs/2310.09497.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023a. Large language models are built-in autoregressive search engines. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2666–2678. Association for Computational Linguistics.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023b. Large language models are built-in autoregressive search engines. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2666–2678. Association for Computational Linguistics.