arXiv:2302.03495v3 [cs.IR] 9 Feb 2023

ChatGPT 能为系统综述文献检索编写出优质的布尔查询吗?

SHUAI WANG, 昆士兰大学，澳大利亚HARRISEN SCELLS,莱比锡大学，德国GUIDO ZUCCON,昆士兰大学，澳大利亚BEVAN KOOPMAN, 澳大利亚联邦科学与工业研究组织，澳大利亚

系统综述是针对高度聚焦的研究问题，对相关文献进行的全面综述。这类综述通常被视为循证医学中证据的最高形式，也是回答医学领域研究问题的关键策略。为了撰写高质量的系统综述，研究人员通常会构建复杂的布尔查询，以检索与综述主题相关的研究。然而，系统综述研究人员往往需要花费很长时间才能构建出高质量的系统综述布尔查询，而且最终得到的查询通常效果不佳。查询质量差可能会导致综述存在偏倚或无效，因为它们未能检索到关键证据；也可能大幅增加综述成本，因为它们检索到了过多无关研究。基于 Transformer 的生成模型近期取得的进展表明，这类模型在有效遵循用户指令并根据指令生成答案方面具有巨大潜力。本文探讨了此类模型中的最新代表——ChatGPT——在为系统综述文献检索生成有效布尔查询方面的效果。通过在该任务的标准测试集上开展多项广泛实验，我们发现 ChatGPT 能够生成带来较高检索精确率的查询，但这是以牺牲召回率为代价的。总的来说，我们的研究展现了 ChatGPT 在为系统综述文献检索生成有效布尔查询方面的潜力。ChatGPT 能够遵循复杂指令并生成精确率较高的查询，因此对于开展系统综述的研究人员而言，它是一种宝贵的工具；尤其是在时间紧迫的快速综述中，通常可以接受以较低召回率换取较高精确率。

ACM 参考格式：
Shuai Wang、Harrisen Scells、Guido Zuccon 和 Bevan Koopman。2023 年。ChatGPT 能否为系统综述文献检索撰写优质布尔查询？1, 1（2023年2月），19 页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

为了构建高质量的系统综述，综述研究人员会检查与研究主题相关的所有证据；其中与研究主题相关的证据将被选出，并进一步评估和综合。为收集与综述相关的所有证据，研究人员会编写布尔查询来检索医学数据库。布尔查询具有可复现性、可解释性，并能过滤掉与研究主题无关的文章 [39], ，从而减少不必要的文献评估工作量。然而，即使对于经验丰富的检索人员而言，构建高质量的布尔查询也颇具挑战，最终生成的查询往往并非最优 [59]。

作者地址：Shuai Wang，昆士兰大学，澳大利亚布里斯班，shuai.wang2@uq.edu.au；Harrisen Scells，莱比锡大学，德国莱比锡，harry.scells@uni-leipzig.de；GuidoZuccon，昆士兰大学，澳大利亚布里斯班，g.zuccon@uq. edu.au；Bevan Koopman，澳大利亚联邦科学与工业研究组织，澳大利亚布里斯班，bevan.koopman@csiro.au。

允许免费为个人或课堂使用而制作本作品全部或部分内容的数字副本或纸质副本，前提是不得为牟利或商业利益而制作或分发副本，且副本必须附有本声明及首页上的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的内容，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布到服务器或分发至列表，则须事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2023 美国计算机协会。XXXX-XXXX/2023/2-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

为帮助研究人员构建更好的系统综述布尔查询，人们开发了自动化布尔查询构建和细化方法 [6, 49, 59, 62–66, 77, 78, 80, 80]。这里所说的查询构建，是指从头创建布尔查询的任务，有时会依赖一篇或多篇“种子”文献，即与该系统综述相关的示例研究。这里所说的细化，是指改进现有布尔查询的任务；例如，通过添加或删除布尔子句和运算符，在保持召回率的同时提高精确率（减少检索到的无关文献）。尽管结果令人鼓舞，且相关方法已被集成到辅助查询创建过程的工具中 [34, 58, 60, 78], ，但自动化技术仍远不能生成高质量的布尔查询。

文本生成模型的最新进展 [21, 35] 在基于任务的问答方面取得了巨大成功，并能根据用户意图提供高质量的回答 [35]。在此类生成模型中，ChatGPT[1],在本文撰写时被认为是表现力最强的文本生成模型，并在自然语言处理任务中取得了当前最先进的效果。本文探讨如何有效利用 ChatGPT 构建高质量的系统综述布尔查询。以下研究问题指导我们对如何使用 ChatGPT 进行调查：

研究问题1： ChatGPT 在构建和优化系统综述布尔查询方面，与当前最先进的方法相比表现如何？**研究问题2：** 用于生成系统综述布尔查询的提示，在多大程度上会影响 ChatGPT 所生成布尔查询的效果？**研究问题3：** 通过多个提示引导 ChatGPT 进行查询构建，并模拟当前最先进的自动布尔查询生成方法的流程，会产生什么效果？**研究问题4：** 使用 ChatGPT 创建系统综述布尔查询有哪些注意事项和潜在挑战？

据我们所知，这是首次对使用 ChatGPT 创建布尔查询的效果进行全面评估的尝试。这包括为 ChatGPT 精心设计提示，并使用系统综述的标准化测试集进行全面评估。

我们发现，ChatGPT 的表现优于当前最先进的查询生成方法。精确率的提升有时会以召回率下降为代价，不过我们表明，通过更好地处理 MeSH术语以及采用滚雪球检索（通过引文网络找到更多研究），这种情况或许可以得到缓解。我们表明，生成一个高质量的提示至关重要。在提示中提供一个布尔查询示例是有益的。与 ChatGPT 进行多轮提示交互优于仅进行一次交互。尽管结果令人鼓舞，但其效果也具有波动性，即效果会因运行次数和提示的不同而变化；尤其是在 MeSH 术语建议方面，ChatGPT 的表现较差。

本研究的发现可以帮助系统综述研究人员使用 ChatGPT 进行系统综述布尔查询构建与优化，并了解其局限性和注意事项。

2 相关工作

2.1 系统综述自动化

构建系统综述需要受过培训的专业人员在多个不同阶段付出人工努力，每个阶段所需的人工工作类型也各不相同。

第1阶段是定义一个具体的研究问题；软件可以帮助定义一个好的查询 [1]。

第 2 阶段是根据研究问题编写布尔查询。查询决定返回哪些结果（以及返回多少），因此其质量会极大影响后续所有阶段。这就是我们特别关注查询

[1]https://openai.com/blog/chatgpt/

生成。人们已开发出用于查询构建 [13, 34, 55], 查询优化 [34, 60]和查询利用 [7, 17]的自动化方法。

第3阶段是筛选通过查询检索到的所有标题和摘要，并人工评估其相关性。此阶段主要的自动化辅助方法是筛选优先级排序 [3–5, 10, 16, 18, 19, 26, 31, 33, 42, 43, 46, 47,61, 68, 69, 81, 82], ，即根据研究与研究问题可能存在的相关性对其进行排序。排序后，自动化方法可以设定“停止标准”，达到该标准后便不再继续筛选，因为此时很可能已找到所有相关文献。筛选过程中通常会采用主动学习方法 [15, 44] ，以帮助对纳入研究进行排序，确定其筛选顺序。

第4阶段是从与综述相关的研究中提取具体细节；数据提取方法可以在此提供帮助 [27, 71]。

第5阶段是将所有证据综合成一份连贯完整的综述文档。为此已开发出综合自动化方法 [12,50, 54, 72, 73] ；其中包括文本主题分析 [75] ，甚至还有文本生成 [51]。

2.2 自动查询构建与优化

查询构建 是根据研究问题，按照一套特定准则推导出布尔查询的过程。该准则介绍了两种公认的系统综述查询构建方法。这些方法描述了构建查询时应遵循的步骤。第一种方法称为概念性方法 [11]。这种方法首先要求从试点检索或事先已知的潜在相关研究中识别若干高层次概念。随后利用这些高层次概念找出同义词和相关关键词。接着对查询进行迭代优化，这一过程由信息专员的专业知识推动。第二种方法称为客观方法 [24, 67]。该方法的第一步是创建一小组潜在相关研究，为后续步骤提供种子，例如像概念方法那样通过试点检索来创建。随后采用多种统计程序，从这些研究中提取术语，并对查询进行初步验证。识别出的术语仍须手动添加到查询中；与概念方法一样，这一过程也依赖信息专员的专业知识。理解本文中这些方法的用途，并不需要了解其具体细节；只需将它们理解为一系列可遵循的步骤，以得出查询即可。

当然，这些方法需要耗费大量时间（因为信息专员必须进行多次试点检索，并花时间验证检索结果） [31, 52] ，而且容易出现人为错误 [56, 57]。为此，Scells 等人 [63] 研究了将这两种查询构建流程自动化的可能性。这一系列研究的主要发现是，将这些流程计算化无法达到人类的效果；然而，对自动生成的查询进行进一步的人工优化，能够显著提升检索效果。

自动查询优化 是基于人工查询优化所展现出的益处而开发的。这类方法会获取人类作者提出的初始查询，并应用一系列转换（添加词项或子句），使查询更加有效 [34, 60]。结合查询可视化 [60] 工具后，这些查询优化工具能够改进初始查询。

从查询构建方法中获得的启示，为使用 ChatGPT 自动构建查询指明了两个明确的方向：一是允许 ChatGPT 按其认为合适的方式生成查询；二是通过提示 ChatGPT 遵循概念性或客观程序中的指示来引导它。我们将第一种方法称为无指导 ，将第二种方法称为引导式。查询优化工作的启示使我们推测，向 ChatGPT 提供一个现有查询并要求其进行优化可能会有所帮助。

2.3 基于 Transformer 架构的生成式语言模型提示工程

提示工程是引导生成式语言模型执行特定任务的过程。从某种意义上说，提示工程可以被视为通过自然语言对生成式语言模型进行编程的一种方式 [53]。通过提示工程引导模型输出的一种流行方法，应用于文本生成图像模型 [37, 48]。这种针对生成式语言模型的任务的“零样本”或“少样本”方法，也在若干自然语言任务中取得了当前最先进的结果 [8]。近来，提示工程也已应用于医学领域的自然语言任务，例如问答 [36, 38]。生成式语言模型在查询构建或优化方面的应用仍未得到充分探索。据我们所知，除本文外，只有另一篇论文发表了使用生成式语言模型进行查询扩展的结果 [14], ，且尚无研究考察过用这种方法创建系统文献综述检索所需的布尔查询。

3 系统综述提示工程

本文探讨使用 ChatGPT 为系统综述文献检索创建查询。ChatGPT 所采用的基本机制是：接收一段文本输入（称为提示），通过模型处理后，输出序列中的下一个词元。此过程会重复多次，以生成对提示的完整回应。ChatGPT 依赖于 Transformer 架构 [76] ，并以海量文本数据进行训练，使其能够学习语言的使用模式。在生成过程中，模型会利用这些已学到的模式，生成与训练时观察到的文本相似的内容。

使用 ChatGPT 并与之交互的一个关键环节，是设计提示来指示模型生成答案。为了指示 ChatGPT 创建用于系统综述文献检索的布尔查询，我们通过反复迭代，设计了一系列复杂程度逐步提高的提示，其中包括依赖示例布尔查询的提示。我们还尝试使用引导查询生成的多轮交互式提示（引导式提示），并借鉴现有的查询制定的客观方法中的见解 [25, 63, 65]。这些提示针对布尔查询生成的两项相关任务而设计：制定任务和优化任务。

3.1 用于查询构建的单一提示

单一查询构建提示是指使用综述标题来指导构建系统综述布尔查询的提示。我们针对布尔查询构建任务设计了五个此类提示；这些提示列于表1中。它们可分为三类：简单、详细和带示例。如表1所示，简单提示是指仅用一句话简要说明 ChatGPT 任务的提示。当用户不擅长构建高质量提示时，这被视为使用 ChatGPT 最常见的方式。另一方面，详细提示是指其中包含背景说明，清楚阐明 ChatGPT 成功完成任务所需内容的提示。带示例提示还包含一个预期的查询构建示例，使 ChatGPT 了解应生成什么样的高质量答案。

两个详细提示“q2”和“q3”的区别在于对背景说明的描述方式；对于带示例提示“q4”和“q5”，提示“q5”还描述了一种用于识别 PICO 要素的布尔查询构建策略。PICO 指的是：患者/问题、干预、对照和结局，通常在进行系统综述检索之前使用，以清晰确定研究问题并生成高质量的系统综述布尔查询 [20]。在提示中加入 PICO 要素提取的考虑是，我们希望帮助 ChatGPT 更好地理解示例查询在逻辑上是如何构建的。

对于每个系统综述主题，我们都会将“{*review_title*}”替换为综述标题。对于包含示例的提示，我们还会将“{*example_review_title*}”替换为示例综述主题的标题，并将“{*example_review_query*}”替换为该示例综述所使用的布尔查询。例如，对于示例标题 “血栓弹力图（*TEG*）和旋转血栓弹力测定法（*ROTEM*）在伴有出血的成年创伤患者创伤诱发性凝血病中的应用”，我们插入以下布尔查询：

	提示 ID	提示
Simple	q1	对于标题为“{review_title}”的系统综述，你能否生成一个系统综述布尔查询，以便在 PubMed 上找到该综述主题的所有纳入研究?
Detailed	q2	你是一名为系统综述开发布尔查询的信息专员。你拥有丰富的拥有开发高效查询以检索医学文献的经验。你的专长是开发能够尽可能减少检索到的无关文献，同时检索到所有相关文献的查询以满足你的信息需求。现在，你需要围绕 {review_title} 开展研究。请构建一个高效的系统综述布尔查询，以尽可能满足你的信息需求。
	q3	请设想你是一位系统综述信息专家；现在，你收到了一项系统综述研究主题，标题为“{review_title}”。你的任务是生成一个高效的系统综述布尔查询，用于在 PubMed 上检索（参考专业编制的查询）；该查询需要尽可能全面，以便检索到所有可纳入该研究主题的相关研究；另一方面，该查询需要尽量减少无关研究的检出数量，以便研究人员可以花更少的时间判断检索到的文档。
With Example	q4	你是一名为系统综述制定布尔查询的信息专员。你拥有丰富的拥有开发高效医学文献检索查询的经验。你的专长是开发能够尽可能减少检出无关文献，同时检出所有相关文献的查询以满足你的信息需求。例如：“{example_review_title}”并生成有效的 PubMed 查询，例如：“{example_review_query}"。现在你有了信息需求，需要针对“{review_title}”开展研究，请生成一个高效的系统综述布尔查询，以满足该信息需求。
	q5	你是一名为系统综述开发布尔查询的信息专员。你拥有丰富经验，擅长开发高效查询以检索医学文献。你的专长是开发查询，尽可能减少检索到的无关文献，并检索出所有相关文献以满足信息需求。专业的信息专员通常会提取 PICO要素，这是构建系统综述布尔查询时的常见做法。PICO 意为患者/问题、干预、对照和结局。PICO 是一种用于开发在开始研究之前提出一个良好的临床研究问题。它是一种助记法，用于描述一个严谨的临床前景问题的四个要素。你可以提出如下信息需求：“{example_review_title}”，并生成有效的 PubMed 查询，例如：“{example_review_query}”。现在你需要围绕“{review_title}”开展研究。首先，从信息需求中提取 PICO要素，并构建一个高效的系统综述布尔查询，以便最大限度地满足你的信息需求。为你的信息需求服务。

表1。单提示查询构建的提示

(血栓弹力图[医学主题词:noexp] OR (thromboelasto*[所有字段]OR thrombelasto*[所有字段] OR ROTEM[所有字段]OR “tem international”[所有字段] OR (thromb*[所有字段] AND elastom*[所有字段]) OR(rotational[所有字段] AND thro mbe- last[所有字段])) OR (Thrombelastogra*[所有字段] OR Thromboelastogra*[所有字段] OR TEG[所有字段] OR haemoscope[所有字段]OR haemonetics[所有字段]OR (thromb*[所有字段] AND elastogra*[所有字段])))

3.2 用于查询优化的单提示

单一查询优化提示是指向 ChatGPT 提供一篇系统综述的标题及其对应的布尔查询，并指示模型修改该查询，以提高检索效果的提示。审稿人可以使用这些提示来改进自己手动编写的查询，

	提示 ID	提示
Simple	q6	对于系统综述的种子布尔查询："{initial_query}"，此查询检索到太多无关文档，且与信息需求“{review_title}”相关的文档太少。请修正此查询，以便检索到更少的无关文档和更多的相关文档。
With Example	q7	对于系统综述的初始布尔查询：“{example_review_initial_query}"，该查询检索到的信息需求相关文献太少，无关文献太多：“{example_review_title}”，因此应将其修正为：“{example_review_refined_query}”。现在，你的任务是针对信息需求“{review_title}”，修正系统综述布尔查询“{initial_query}”，使其检索到更少的无关文献和更多的相关文献。

表2. 用于单提示查询优化的提示

也可以将其整合到更复杂的自动化流程中。我们在表2中展示了设计的查询优化提示。

我们将用于查询优化的提示分为简单和带示例两类。与单提示词查询构建类似，简单提示“q6”表示该提示仅通过简要描述任务来构建。对于带示例提示“q7”，我们加入一个示例，向 ChatGPT 说明成功优化查询意味着什么。要求 ChatGPT 优化查询时，我们用希望 ChatGPT 优化的初始布尔查询替换“{$initial_query$}”。对于需要示例的提示，我们还会将 {$example_review_title$} 替换为示例综述主题的标题；将“{$example_review_initial_query$}”替换为该示例主题的初始查询；并将“{$example_review_refined_query$}”替换为该示例主题中优化后的最终查询。

我们研究了两类示例的效果：（1）一个高质量系统综述示例（HQE）；（2）一个与查询主题相似或相关的示例（RE）。为了找到相关示例，我们使用 monoBERT 架构，从测试集 [45] 中找出最接近的示例。具体而言，我们将查询主题的综述标题与一个潜在示例综述的标题拼接起来，并将其输入 PubMedBERT——一种使用 PubMed摘要预训练的领域专用 BERT 模型 [23]。我们获得一个最终分类分数，用以衡量潜在示例主题与查询主题的相关性；创建提示时，我们选择排名最高的示例并将其纳入提示。

3.3 查询构建的引导式提示

我们设计了一个多步骤提示，其逻辑遵循当前最先进的自动化查询构建方法之一，即目标方法 [63]。我们在表3中通过示例展示了设计的引导式提示。具体而言，我们采用四步流程来生成布尔查询：第一步，使用一项种子研究确定布尔查询术语；第二步，将这些术语分为四类：与健康状况相关的术语（A）、与治疗相关的术语（B）、与研究设计类型相关的术语（C）或其他术语（N/A）。第三步，将同一类别中的术语通过“OR”组合，并用“AND”组合最终查询。最后，请 ChatGPT 通过添加更多术语（如医学主题词）来优化查询。借助引导式提示，我们可以更好地控制 ChatGPT 构建布尔查询时使用的术语，还能将复杂任务拆分为多个更易处理且有效的子任务。

4 实验设置

我们的实验使用两个数据集：CLEF 技术辅助综述（TAR）数据集 [28–30]，以及包含种子研究的系统综述集合（种子集合） [79]。CLEF TAR 数据集于 2017年 至 2019年 每年发布，作为验证数据集，用于改进系统综述的布尔查询构建和筛选。我们使用 CLEF TAR 2017 和 2018，因为两者都只包含诊断性试验准确性系统综述。由于 CLEF TAR 2017 中的综述是 CLEF TAR 2018 数据集的子集，因此在去重后，我们获得了 72 个综述主题

Step	提示	示例回答
步骤1	严格遵循我的指示，制定一个高效的布尔查询，用于医学系统综述文献检索。不要解释或详细说明。只按我的要求作答。首先，根据以下陈述和一项相关研究的文本，请找出 50 个相关术语或短语。你找出的术语应当用于检索更多相关研究，因此请谨慎选择术语，确保它们范围不要太宽泛。列表中不得有重复项。 陈述：“尸检研究中的分化型甲状腺癌患病率——数十年来的病例：荟萃分析” 文本：在274例未选尸检中，病理科在何塞·华金·阿吉雷医院发现了10例甲状腺隐匿性癌，时间为1980年12月至1983年3月。这是最低的发病率在最新的已发表病例系列中。最小的乳头状微小癌最常表现为无包膜、无硬化的肿瘤中没有炎性浸润或纤维化，这提示它们可能代表早期发展阶段。我们的数据表明，乳头状微小癌的频率相对较高，且在外科手术和尸检系列中相近，但并未显示临床进展的危险因素。	1. 分化型甲状腺癌 2. 患病率 3. 尸检研究 4. 荟萃分析 5. 隐匿性癌 6. 甲状腺 7. 发病率 8. 环境因素 9. 病因学 10. 低风险 ... 50. 参数。
步骤 2	对于你在步骤 1 中创建的列表中的每个条目，将其划分为以下三类之一类别：与健康状况相关的术语（A）、与治疗相关的术语——（B），与研究设计类型相关的术语（C）。当某项不符合这些类别之一时，将其标记为（N/A）。每项都需要归入（A）、（B）、（C）或（N/A）。	1. （A）分化型甲状腺癌 2. （N/A）患病率 3. （C）尸检研究 4. (C) 荟萃分析 5. (A) 隐匿性癌 6. (A) 甲状腺 7. (N/A) 发病率 8. (N/A) 环境因素 9. (N/A) 病因学 10.（不适用）低风险 ... 50.（不适用）参数。
步骤 3	使用你在步骤 2 中创建的分类列表，创建一个布尔查询，并将其提交至 PubMed。将每个类别归为一组项目。例如：((itemA1[标题/摘要] OR itemA2[标题/摘要]) AND (itemB1[标题/摘要] OR itemB2[标题/摘要]) AND (itemC1[标题/摘要] OR itemC2[标题/摘要] 或 itemC3[标题/摘要]))	(((分化型[标题/摘要] OR 甲状腺[标题/摘要] OR 癌[标题/摘要] OR 乳头状[标题/摘要] OR microcarcinoma[标题/摘要]) AND (癌症[标题/摘要] OR 发病率[标题/摘要] OR 病因学[标题/摘要] OR 风险[标题/摘要] OR 性别[标题/摘要] OR hormonal[标题/摘要] OR 结节性的[标题/摘要] OR goiter[标题/摘要] OR 桥本氏[标题/摘要] OR 恶性肿瘤[标题/摘要] OR 伴随的[标题/摘要] OR 肿瘤[标题/摘要] OR 浸润[标题/摘要] OR fibrosis[标题/摘要] OR 发展[标题/摘要] OR 频率[标题/摘要])) AND (尸检[标题/摘要] OR 外科手术[标题/摘要] OR 材料[标题/摘要] OR 系列[标题/摘要] OR 样本[标题/摘要] OR 病例[标题/摘要])).
步骤 4	运用你的专业知识来优化查询，使其尽可能检索到更多相关文献，同时尽量减少文档总数。 检索到的文献。此外，在适当情况下将相关医学主题词添加到查询中，例如，MeSH术语[MeSH]。保留查询的总体结构，但查询的每个主要子句都应对应一个 PICO 要素，。最终查询仍需能在 PubMed 上执行，因此应当是有效查询。	(((分化型甲状腺癌[MeSH] OR "分化型甲状-腺"[所有字段]或“甲状腺癌”[所有字段] 或“乳头状微小癌”[所有字段]) 与 (患病率[所有字段] OR 发病率[医学主题词] 或“病因学”[所有字段] 或“危险因素”[所有字段] 或性别[所有字段] 或激素因素[所有字段] 或“结节性甲状腺肿”[所有字段] 或“桥本甲状腺炎”[医学主题词]或恶性-nancy[医学主题词] 或“伴发病变”[所有字段] 或肿瘤[所有字段] 或浸润[所有字段] 或纤维化[所有字段] 或“早期发展阶段”[所有字段]或频率[所有字段])) 与 (尸检[医学主题词] 或外科手术[所有字段] 或材料[所有字段] 或系列[所有字段]或样本[所有字段] 或病例[所有字段]))

表3. 用于查询制定的引导式提示示例。

。种子集合包含 40 个系统综述主题及真实的种子研究，这些研究可在布尔查询构建过程中为系统综述研究人员提供指导。两个集合均包含系统综述标题、布尔查询和相关评估。

		提示	精确率	F1	F3	召回率
CLEF	基线	原始	0.0207*	0.0290*	0.0481*	0.8317*
		概念性	0.0015*	0.0027*	0.0101*	0.6997*
		客观	0.0002*	0.0005*	0.0023*	**0.9128***
	简单	q1	0.0543	0.0500	0.0590	0.1293*
	详细	q2	**0.1166**	0.0654	0.0696	0.1310*
		q3	0.0844	0.0443	0.0497*	0.1175*
	带示例	q4	0.0752	0.0642	**0.0847**	0.5035
		q5	0.0958	**0.0717**	0.0844	0.3335*
SC	基线	原始	0.0367	**0.0651***	**0.1099***	**0.7366***
		概念性	0.0018			0.4138
		客观	0.0057			0.5192
	简单	q1	0.0501	0.0274	0.0298	0.0528
	详细	q2	**0.0983**	0.0310	0.0278	0.0394*
		q3	0.0730	0.0329	0.0329	0.0519
	带示例	q4	0.0283	0.0274	0.0374	0.1290
		q5	0.0188	0.0193	0.0271	0.0785

表4。单提示词查询构建结果。 *CLEF* 表示 CLEF TAR 集合， *SC* 表示种子集合。q4 与所有其他方法之间的统计显著差异（采用 Bonferroni 校正的 Student 双尾配对 t 检验，$p < 0.05$）以 *表示。

为了生成提示，我们将表1、2和3所示的设计提示中的字段替换为集合中的特定综述主题信息。对于包含示例的提示，我们选取 CLEF TAR 集合中的主题 CD010438 作为高质量示例主题，为两个集合生成查询，因为该查询效果较好，且结构简单，便于 ChatGPT 遵循。

由于无法获得原始集合中报告的由目标方法和概念方法生成的种子集合查询 [79]，因此我们没有对种子集合运行查询优化。引导式提示方法依赖于目标方法，而目标方法需要至少一个种子研究才能开始构建查询。因此，我们没有在 CLEF TAR 上运行引导式提示，因为其中不包含种子研究。

从 ChatGPT 获得生成的布尔查询后，对于综述主题，任何生成格式错误布尔查询的查询都会被删除并重新生成。

为评估生成的布尔查询，我们使用 PubMed 的 Entrez API 执行这些查询，以获取检索到的 PubMed ID [9]。随后，我们使用精确率、F 值和召回率等基于集合的度量，并根据集合中的摘要级相关性评估来评估检索到的 PubMed ID。

5 主要结果

5.1 单提示查询构建

我们在表4中展示了单提示词查询构建的结果。结果表明，与当前最先进的自动查询构建方法相比，ChatGPT 生成的查询通常具有更高的精确率，但代价是召回率较低。就 F 值而言（同时体现精确率和召回率），在 CLEF 集合上，ChatGPT 生成的查询比当前最先进的查询和原始作者编写的查询都更有效。然而，在种子集合上，其效果较差。在系统综述中，检索通常

		精确率	F1	F3	召回率
CLEF	q4-高质量示例	0.0751	0.0642	0.0872	0.5035
	q4-相关示例	0.1105(+47.1%)	0.0909(+41.6%)	0.1144(+31.2%)	0.4183(−38.1%)
SC	q4-高质量示例	0.0283	0.0274	0.0374	0.129
	q4-相关示例	0.0351(+24.0%)	0.0140(−48.9%)	0.0139(−62.8%)	0.0161*(−87.6%)

表5。使用不同类型示例时，单提示查询生成提示“q4”的结果对照。 *CLEF*表示 CLEF TAR 集合， *SC* 表示种子集合；对于每个集合，使用两种类型的示例， $q4-HQE$表示使用一个高质量示例，而 $q4-RE$ 表示使用相关查询作为示例。两种示例类型之间的统计显著差异（$p < 0.05$）用 *表示。

需要高召回率，以确保找到与研究主题相关的所有证据。所有由 ChatGPT 生成的查询，其召回率都低于基线方法，这表明 ChatGPT 生成的查询可能不适合高召回率检索，而更适合时间有限的情况，例如快速综述[40]。

使用不同的简单和详细提示（q1–3）对效果的影响很小。对于 CLEF，q2 在精确率方面显著优于其他提示；除此之外，提示类型并未产生显著影响。不过，我们发现，将高质量系统综述主题作为示例纳入提示，效果显著优于不含示例的提示，表现为 F_1、F_3 和召回率始终更高。在比较两种包含示例的提示的效果时，我们发现，先要求 ChatGPT 生成 PICO 要素，再生成最终的布尔查询，得到的布尔查询召回率较低，但精确率较高。总体而言，我们的研究结果表明，在提示中加入高质量的系统综述查询示例至关重要，而任务描述的详细程度可能不会产生显著影响。

接下来，为了评估查询构建示例对使用 ChatGPT 生成的布尔查询效果的影响，我们进一步按照第3.1节所述，测试使用不同类型示例时的效果。

表5比较了提示中采用最相关主题生成的查询与采用高质量示例生成的查询的效果。将相关主题用作示例，可能会得到精确率较高但召回率较低的查询。

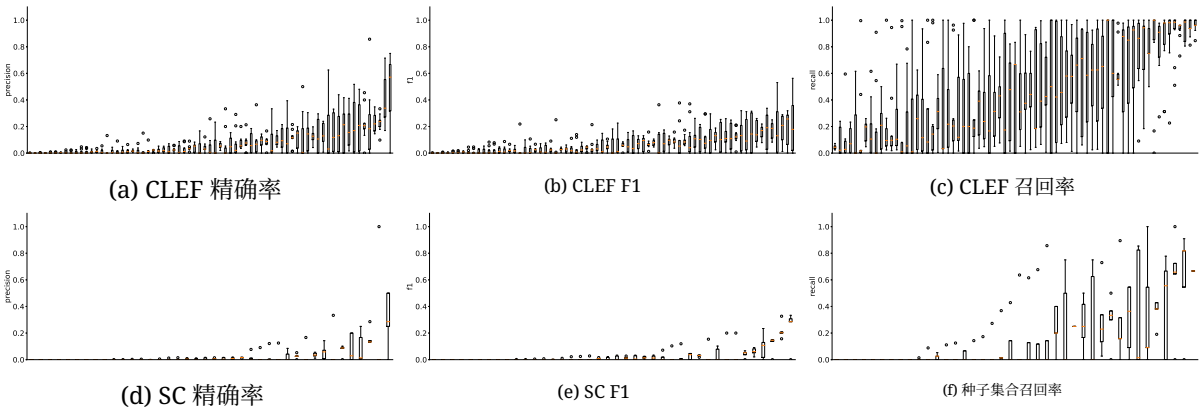
对于相同的提示，ChatGPT 会生成不同的回答。为研究效果的变异性，我们选取效果更好的提示 q4，并将其重新运行 10 次。变异性如图1所示。召回率的波动幅度大于精确率和 F 度量：召回率的方差为其均值的 12%；精确率为 7.1%，F-1值为 6.6%，F-3值为 7.2%。我们的结果表明，使用相同提示生成的查询，主要区别在于从纳入的研究中获取更多相关文档的能力（召回率）。

5.2 单提示查询优化

表 6 所示的单提示查询优化结果表明，ChatGPT 能够提升系统综述布尔查询的效果。具体而言，使用 ChatGPT 优化查询可提高精确率和 F 值，但会降低召回率。因此，首先创建一个高召回率的种子查询至关重要，随后再使用 ChatGPT 优化该查询，以获得效果出色的布尔查询。

此外，我们的结果表明，使用目标方法自动构建的种子查询时，效果提升最大。这种方法还能在所有方法中得到召回率最高的查询。经 ChatGPT 优化后，召回率仅下降 11%，而精确率、F-1值和 F-3值则显著提升。

变异性 再次通过将表现最佳的方法——优化后的目标方法——运行 10 次进行研究。我们展示了 e
.....的变异性 图2中的查询优化。与查询构建相比，查询优化的差异较小 n



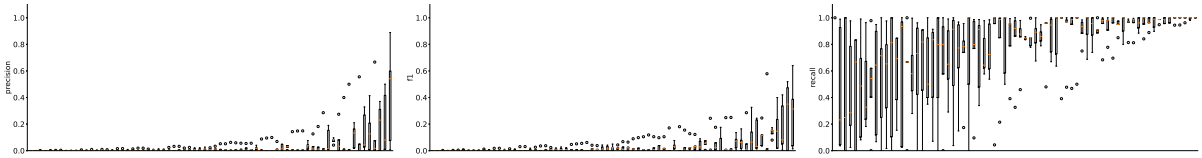
(a) CLEF 精确率 (b) CLEF F1 (c) CLEF 召回率

(d) SC 精确率 (e) SC F1 (f) 种子集合召回率

图1. 按主题划分的 单提示词查询构建中10次迭代运行效果的变异性箱线图。 *CLEF*表示 CLEF TAR 集合， *SC* 表示种子集合。

提示	精确率	F1	F3	召回率
原始	0.0207	0.0290	0.0481	0.8317
q6-原始	0.0795*	0.0597*	0.0802*	0.5060*
概念性	0.0014	0.0027	0.0100	0.6996
q7-概念性	0.0022	0.0039	0.0069	0.2699*
客观	0.0002	0.0005	0.0023	**0.9128**
q7-客观	0.0460*	0.0471*	0.0652*	0.8115*
q4	0.0751	0.0642	0.0872	0.5035
q7-q4	**0.1162**	**0.0772**	**0.0921**	0.3179*

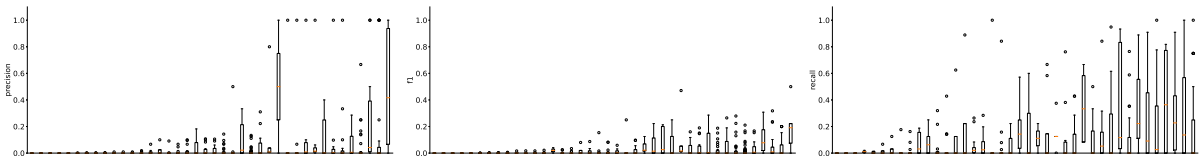
表6. CLEF TAR 集合上的单提示查询优化结果表。对于细化方法“q6-Original”，“q6”表示用于生成细化查询的提示；“原始”表示用于让 ChatGPT 进行细化的种子查询。对于每种查询优化方法，细化提示与种子查询之间具有统计显著差异（$p < 0.05$）的情况以 * 表示。



(a) 精确率 (b) F1 (c) 召回率

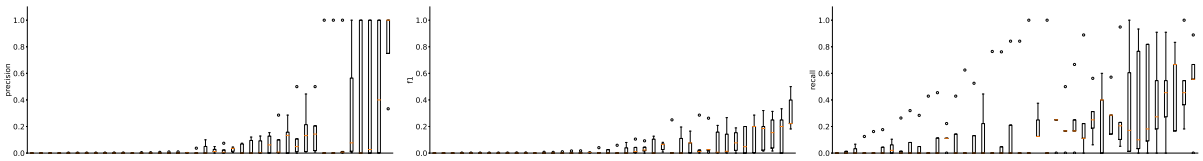
图2. To按主题划分的单提示查询优化中 10 次迭代运行效果的变异性箱线图 .

（图1）。这可以理解，因为种子查询已经提供了查询结构，而查询构建则必须根据综述标题从头开始。



(a) 精确率 (b) F1 (c) 召回率

图3. 按主题划分 用于评估不同种子研究在引导式提示查询构建中的效果的变异性箱线图 n.



(a) 精确率 (b) F1 (c) 召回率

图4. 使用相同种子研究进行引导式提示查询构建的10次迭代运行效果按主题变化的箱线图。

提示	精确率	F1	F3	召回率
q4	0.0284	0.0274	0.0374	0.1290
引导式	0.0993(+249.6%)	0.0492(+79.6%)	0.0565(+51.1%)	0.5171(+301.6%)*

表7. 种子集合上引导式提示查询构建的结果表，与单提示查询生成“q4”相比；引导式提示与单提示之间的统计显著差异（$p < 0.05$）以 *标示。

5.3 引导式提示查询构造

表7中引导式提示查询构建的结果表明，若使用经过精心挑选的种子研究，由引导式提示生成的查询比单提示查询更有效。

然而，与先前关于单提示查询生成和优化的实验一样，查询效果在不同运行之间存在显著差异；查询效果也取决于所使用的种子研究。图3展示了使用不同种子研究生成查询时查询效果的变异性；图4则展示了使用同一种子研究时效果的变异性，该种子研究选自第一次运行中效果最佳的种子研究。

从变异性图中可以看出，使用引导式提示生成查询时，不同种子研究之间的结果并不稳定。此外，即使使用同一种子研究进行多次运行，结果仍存在高度变异性。对于某些主题，精确率和召回率的范围可能从0到1，尤其是在平均效果较高时。

6 查询失败分析

我们试图揭示效果不佳的查询所具有的特征，表面上则是为了找出改进查询生成的方法。为此，我们按照以下步骤选出一组成功和失败的查询：(1) 选出查询构建（q4）和查询优化（q7-objective）的最佳方法。

(2) 从每项任务的 10 次迭代运行中，我们提取一个“预言机”结果，并以召回率作为效果指标。（若出现平局，则以精确率打破平局。）
(3) 接下来，我们将预言结果有效性与原始查询的效果进行比较。成功主题是指预言结果的精确率和召回率均高于原始查询的主题；失败主题是指预言结果的精确率和召回率均低于原始查询的主题。上述选择成功主题和失败主题的直觉依据是：更高的召回率通常意味着查询检索到更多文档，从而导致更低的精确率。
我们仅使用 CLEF 数据集，因为种子集合中的主题不一定是系统综述；例如范围综述、快速综述等。

采用上述方法，我们确定了查询构建的七个成功主题[2] 和六个失败主题 [3] 。对于查询优化，我们确定了三个成功主题 [4]，以及十一个用于查询优化的主题 [5]。

通过将失败查询与其他查询进行比较，我们总结出三个主要发现：
(1) 表现不佳的 ChatGPT 生成查询往往会检索大量结果：在比较查询构建中的成功查询和失败查询时，成功查询的比率中位数为 0.30，而失败查询的比率中位数为 10.28。查询优化也有类似情况，成功查询和失败查询的比率中位数分别为 1.39 和 2.74。这意味着失败查询可能具有 OR 子句数量较多的特征；因此，或许可以检测出此类查询并引入缓解策略。

(2) 我们还注意到，一些 ChatGPT 查询包含许多错误的 MeSH 术语。对于查询构建效果最佳的提示，平均生成了三个 MeSH 术语，其中 55% 不在 MeSH 词汇表中。另一方面，对于查询优化提示，每个查询平均只生成 1.5 个 MeSH 术语，其中 66% 不在 MeSH 词汇表中。错误生成的 MeSH 术语所占比例与查询效果之间没有明显相关性。不过，我们也承认，这并不意味着某些错误的 MeSH 术语不会对效果产生重大负面影响。还需要进一步实验（例如 MeSH 术语纠正 [77, 80]），以检验这些 MeSH 术语会如何影响查询效果，以及纠正 MeSH 术语后效果会如何变化。

(3) 相关性判断基于原始查询进行，因此可能会偏向原始查询。在所有失败的查询中，检索文献集中未判定文献的数量占检索文献总数的 94.8% 以上，远高于原始查询的 63.7%；另一方面，其他主题中生成查询与原始查询所检索文献里的未判定文献比例相近，分别为 56.2% 和 55.4%。失败的查询可能只是因为效果不佳而检索到了更多未判定文献。但也有一种可能：它们实际上找到了与主题相关、却从未经过审阅者评估的文档簇。尽管这只是推测，但确实引出了一些改进此类查询的思路。如果大量不相关文献中确实存在相关文档簇，那么自动分类技术或许可以帮助发现这些文档簇 [2, 32, 41, 70, 74]。具体而言，可以使用主动学习 [15, 44]，有选择地评估不同文档簇中的文献，从而逐步锁定相关文献。

7 研究结果摘要

鉴于第5节报告的实证结果，我们根据所设定的研究问题，总结本研究的主要发现。

[2]主题 CD008759、008686 3 009323、007427、010023、010213、011515、CD CD CD CD CD CD 主题 CD010705、CD009591、CD010680、CD010502、CD010296 和 CD011126。[4]主题 CD010023、007427 5 010213、CD CD 主题 CD010657、CD012165、CD009135、CD011926、CD008643、CD011134、CD012083、CD008782、CD009579、CD008691、CD012010。

RQ1： 与当前最先进的方法相比，*ChatGPT* 在制定和优化系统综述布尔查询方面表现如何?

就自动查询构建而言，我们的结果表明，与当前最先进的自动布尔查询生成方法相比，使用ChatGPT在精确率方面表现良好，但代价是召回率较低。

就自动查询优化而言，我们的结果表明，ChatGPT 能够有效优化由自动布尔查询构建的前一步骤生成的现有布尔查询。特别是，基于 ChatGPT 的优化对于使用客观布尔查询构建方法生成的查询，能够带来最显著的效果提升。

在这两项任务中，我们都观察到精确率有所提高，即查询检索到的无关文档减少。这些改进是以召回率下降为代价的（不过，在最佳查询优化设置下，这种下降幅度仅为边际程度）。对于传统系统综述而言，高召回率通常是一项关键要求，因此 ChatGPT 可能并不适合这些场景。不过，我们注意到，在人工创建的布尔查询中，通常会使用医学主题词来提高召回率；但通过 ChatGPT 获得的查询包含大量错误的医学主题词（见第 6 节）。增加一个用于纠正 ChatGPT 生成的错误医学主题词的后处理步骤，或许可以缓解我们在结果中观察到的召回率损失。为此，可以考虑调整现有的 MeSH 主题词推荐方法 [77, 80],，这些方法也可用于进一步优化 ChatGPT 生成的查询，通过添加更多有效的医学主题词来补充模型可能未能识别出的术语。我们还注意到，在系统综述过程中，滚雪球检索旨在进一步提高召回率。滚雪球检索是指对检索文档中所引用参考文献进行递归分析 [22] ——实际上，这一过程会将布尔查询未检索到的研究补充到待评审研究集合中。因此，使用 ChatGPT 创建查询时所观察到的召回率损失，或许可以通过对已识别的相关研究进行滚雪球检索来弥补。

研究问题2： 用于生成系统综述布尔查询的提示，在多大程度上会影响 *ChatGPT* 所生成布尔查询的效果?

我们的结果表明，所用提示的类型会显著影响 ChatGPT 生成的查询效果。我们将在下一个研究问题中进一步探讨引导式提示的影响，而此处仅关注单提示。

在这种情况下，我们发现，在提示中加入布尔查询示例，往往比使用简单提示生成的查询更好（表 4）。这一点在 CLEF 数据集上尤其明显；尽管在种子集合中并非如此，但该集合涵盖的主题数量有限，因此结果的普适性可能较低。值得注意的是，加入查询示例所带来的召回率提升幅度较大（种子集合也有这种情况），从而在一定程度上弥补了 ChatGPT 生成的查询普遍存在的低召回率。我们还发现，示例查询的性质会影响结果。具体而言，与所要创建新查询的系统综述主题在语义上相近的示例查询，能带来更高的精确率。另一方面，在提示中使用质量较高但可能无关的查询，则能带来更高的召回率。

我们还观察到，在提示中加入生成 PICO 元素的指令会导致召回率大幅下降，而且只有在 CLEF 集合上才能提高精确率。

研究问题3： 通过多个提示引导 *ChatGPT* 构建查询，并模仿当前最先进的自动布尔查询生成方法，这会产生什么效果?

我们借鉴现有客观查询生成方法 [63, 65]，设计了一系列提示，逐步引导 ChatGPT 生成布尔查询。结果表明，引导式提示比单提示策略效果更好：精确率和召回率均有所提高。不过，也有一些需要留意的问题——我们将在下文讨论。**研究问题4**

：*What are th*这些问题*ts and potential chall*包括 *f using Cha*tGPT t 生成 *t*系统*temati*综述*iew Bool*布尔查询*ies?*

我们的结果指出，如果依赖 ChatGPT 创建用于系统综述文献检索的布尔查询，实践者应警惕两项主要注意事项：（1）MeSH术语错误；（2）多次请求之间的查询效果存在较大变异性。

我们在分析 RQ1 时已经提到了第一个注意事项。ChatGPT 生成的医学主题词中，55% 及以上实际上并不在 MeSH 词汇表中，因此是错误的。这些错误医学主题词的影响目前仍不明确，因为现有的自动医学主题词推荐方法能否改进，以纠正 ChatGPT 生成的医学主题词，尚不清楚 [77, 80]。

第二个注意事项涉及图 1–4 所示的结果。这些结果表明，如果针对同一个提示多次使用 ChatGPT 生成内容，生成的查询会有所不同，其效果也可能存在很大差异。给定相同的提示，ChatGPT 的输出具有非确定性[6]，这表明它可能使用了 top-k采样或类似技术进行文本生成。top-k采样会针对给定输入生成最有可能出现的 k下一个词元，然后随机选择其中一个词元，作为生成过程中的下一步。尽管如此，无论具体采用何种文本生成技术，ChatGPT 的非确定性，以及这一特性对所生成布尔查询的效果造成的显著影响，都会对本文所研究方法的推广应用构成威胁。事实上，用户通常无法或只能有限地了解 ChatGPT 所生成查询的效果，因此可能无法从针对同一提示生成的多个查询中识别出高产查询。

8 使用ChatGPT的局限性

在我们的实验中，我们将 ChatGPT 视为一个黑箱应用，并测试其生成系统综述布尔查询的效果。本文的一项重要局限是，ChatGPT 内部工作原理的细节并未公开，包括模型架构以及训练数据所包含的内容。因此，我们承认，在模型训练期间，ChatGPT 可能已经接触过测试集中各系统综述主题的原始查询，这可能会影响 ChatGPT 生成的查询的效果。此外，我们不清楚交互历史（超出对话的即时上下文；服务提供商 OpenAI 表示，即时上下文约为 3000 个单词或 4000 个词元）是否会影响生成。虽然在针对新查询主题开展实验时，我们会在 ChatGPT 中开启新的对话，但我们不知道服务提供商会为用户保留哪些信息，也不清楚他们是否会利用用户以往的对话来个性化生成结果。

第三，我们的实验在 2023 年 1 月上旬的两周时间内进行。我们不确定这段时间内是否对模型进行了修改。OpenAI 披露，我们使用的模型是 1月9日版本；在 OpenAI 将模型更新至 1月30日版本后，我们没有再进行实验。然而，我们无法控制 OpenAI 可能对模型进行但未披露的细微工程调整或模型更新。

9 结论

本文使用 ChatGPT 为系统综述构建和优化布尔查询。我们设计了一套全面的提示，用于在 100 多个系统综述主题上研究这些任务。与当前最先进的自动查询构建方法生成的查询相比，ChatGPT 生成的查询精确率更高，但召回率更低。我们还表明，通过使用引导式提示，ChatGPT 能够生成效果更好的布尔查询。我们的结果有一个值得注意之处：即使使用相同的提示，ChatGPT 也会生成不同的查询，且这些查询的效果各异。在将 ChatGPT 用于

[6]我们通过以下方式对此进行了验证：在不同的“聊天”中发送提示并进行其他互动，以避免 ChatGPT 使用之前的 us
输入和输出作为 当前交互的输入部分。

实际的系统综述文献检索中生成布尔查询之前，需要先解决这一问题，因为可复现性至关重要。

总之，本文作出了以下贡献：

- 据我们所知，本文是首篇研究 ChatGPT 生成的、用于系统综述文献检索的布尔查询效果的论文。
- 几种为 ChatGP 生成提示的策略。
- 对大量系统综述主题和提示进行广泛评估，有助于了解不同提示如何影响通过 ChatGPT 生成的系统综述布尔查询的效果。
- 查询失败分析为未来研究提供了可能的方向，包括利用 ChatGPT 生成更好的查询，或对这些查询的结果进行后处理。总体而言，基于 Transformer 的生成模型（如 ChatGPT）能否构建所有布尔查询，目前仍不清楚。本文展示了此类模型的潜力，但也指出了围绕变异性、稳健性和可复现性的诸多问题。无论结果如何，我们坚信这是一个前景广阔、令人振奋的未来研究领域。

参考文献

[1] 2019。Review Manager Web（RevMan Web）。[2] JJ García Adeva、JM Pikatza Atxa、M Ubeda Carrillo 和 E Ansuategi Zengotitabengoa。2014年。《用于支持医学系统综述的自动文本分类》。《Expert Systems with Applications》41卷第4期（2014年），1498–1508页。[3] Amal Alharbi、William Briggs 和 Mark Stevenson。2018年。《系统综述的研究检索与排序：谢菲尔德大学在 CLEF eHealth 2018 任务2中的方法》。载于《CEUR研讨会论文集：CLEF 2018评测论坛会议与实验室工作论文》，第2125卷。CEUR研讨会论文集。[4] Amal Alharbi 和 Mark Stevenson。2017年。《对摘要排序以识别系统综述的相关证据：谢菲尔德大学在 CLEF eHealth 2017 任务2中的方法》。载于《CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文》。[5] Antonios Anagnostou、Athanasios Lagopoulos、Grigorios Tsoumakas 和 Ioannis P Vlahavas。2017年。《结合综述间学习排序与综述内增量训练，用于系统综述的标题与摘要筛选》。载于《CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文》。[6] Maisie Badami、Boualem Benatallah 和 Marcos Baez。2022年。《系统文献综述搜索查询优化流水线：增量丰富与适应》。载于《国际高级信息系统工程会议》。施普林格，129–146页。[7] Florian Boudin、Jian-Yun Nie 和 Martin Dawes。2010年。《利用文档与 PICO 结构进行临床信息检索》。载于《人类语言技术：北美计算语言学协会分会2010年年会论文集》，822–830页。[8] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D. Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry 和 Amanda Askell。2020年。《语言模型是少样本学习器》。《神经信息处理系统进展》33（2020年），1877–1901页。[9] Kathi Canese 和 Sarah Weis。2013年。《PubMed：书目数据库》。《NCBI手册》2卷第1期（2013年）。[10] Jiayi Chen、Su Chen、Yang Song、Hongyu Liu、Yueyao Wang、Qinmin Hu、Liang He 和 Yan Yang。2017年。《华东师范大学参加2017年电子健康任务2：实证医学技术辅助综述》。载于《CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文》。[11] Justin Clark。2013年。《系统综述方法》。载于《临床流行病学方法》，Gail M. Williams、Suhail A. R. Doi（编）。施普林格。[12] Justin Clark、Paul Glasziou、Chris Del Mar、Alexandra Bannach-Brown、Paulina Stehlik 和 Anna Mae Scott。2020年。《借助自动化工具在两周内完成一项完整的系统综述：案例研究》。《临床流行病学杂志》121卷（2020年），81–90页。[13] Justin Michael Clark、Sharon Sanders、Matthew Carter、David Honeyman、Gina Cleo、Yvonne Auld、Debbie Booth、Patrick Condron、Christine Dalais、Sarah Bateup 等。2020年。《利用 Polyglot 搜索翻译器改进检索策略翻译：一项随机对照试验》。《医学图书馆协会杂志》108卷第2期（2020年），195页。[14] Vincent Claveau。2021年。《用于信息检索中查询扩展的神经文本生成》。载于《IEEE/WIC/ACM国际网络智能会议》。ACM，澳大利亚维多利亚州埃森登，202 209 10 1145 3486622 3493957 –。https://doi.org/。/。[15] A.M. Cohen、W.R. Hersh、K. Peterson 和 P.Y. Yen。2006。。《利用自动引文分类减少系统综述准备工作量》。《JAMIA》13卷第2期（2006年），206–219页。[16] Aaron M Cohen 和 Neil R Smalheiser。2018年。《UIC/OHSU 参加 CLEF 2018 任务2：利用出版物类型聚类相似性度量进行诊断试验准确性排序》。载于《CEUR研讨会论文集：CLEF 2018评测论坛会议与实验室工作论文》，第2125卷。[17] Dina Demner-Fushman 和 Jimmy Lin。2007年。《运用基于知识和统计技术回答临床问题》。《计算语言学》33卷第1期（2007年），63–103页。[18] Giorgio Maria Di Nunzio、Federica Beghini、Federica Vezzani 和 Geneviève Henrot。2017年。《系统综述中查询方面重写的交互式二维方法：IMS Unipd 参加 CLEF eHealth 任务2》。载于《CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文》。[19] Giorgio Maria Di Nunzio、Giacomo Ciuffreda 和 Federica Vezzani。2018年。《系统综述的交互式采样：IMS Unipd 参加 CLEF 2018 eHealth 任务2》。载于《CEUR研讨会论文集：CLEF 2018评测论坛会议与实验室工作论文》。[20] Mette Brandt Eriksen 和 Tove Faber Frandsen。2018年。《患者、干预、对照、结局（PICO）作为搜索策略工具对文献检索质量的影响：一项系统综述》。《医学图书馆协会杂志：JMLA》106卷第4期（2018年），420页。[21] Roberto Gozalo-Brizuela 和 Eduardo C Garrido-Merchan。2023年。《ChatGPT 并非你所需的一切：大型生成式人工智能模型的最新进展综述》。arXiv预印本 arXiv:2301.04655（2023年）。[22] Trisha Greenhalgh 和 Richard Peacock。2005年。《复杂证据系统综述中检索方法的效果与效率：一手来源审计》。《BMJ》331卷第7524期（2005年），1064–1065页。[23] Yu Gu、Robert Tinn、Hao Cheng、Michael Lucas、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。2020年。《面向生物医学自然语言处理的领域特定语言模型预训练》。arXiv:arXiv:2007.15779

[24] Elke Hausner、Charlotte Guddat、Tatjana Hermanns、Ulrike Lampert 和 Siw Waffenschmidt。2015年。系统综述检索策略的制定：验证研究表明，客观方法并不劣于其他方法。《临床流行病学杂志》68，2（2015年），191–199。[25] Elke Hausner、Siw Waffenschmidt、Thomas Kaiser 和 Michael Simon。2012年。常规制定客观导出的检索策略。《系统综述》1，1（2012年），19。[26] Noah Hollmann 和 Carsten Eickhoff。2017年。面向高召回率文档检索的排序与基于反馈的停止策略。。载于 CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文。[27] William Hsu、William Speier 和 Ricky K Taira。2012年。报告统计分析的自动提取：实现临床试验文献的逻辑表示。载于 AMIA年度研讨会论文集，第1卷（2012年）。350。[28] E. Kanoulas、D. Li、L. Azzopardi 和 R. Spijker。2017年。CLEF 2017实证医学技术辅助综述概述。载于 CLEF'17。[29] Evangelos Kanoulas、Dan Li、Leif Azzopardi 和 Rene Spijker。2019年。CLEF 2019实证医学技术辅助综述概述。载于CEUR研讨会论文集，第2380卷。[30] Evangelos Kanoulas、Rene Spijker、Dan Li 和 Leif Azzopardi。2018年。CLEF 2018实证医学技术辅助综述概述。载于 CLEF 2018评测实验室与研讨会：在线工作论文，CEUR-WS。[31] Sarvnaz Karimi、Justin Zobel、Stefan Pohl 和 Falk Scholer。2009年。生物医学系统检索中的高召回率挑战。载于第3届生物信息学数据与文本挖掘国际研讨会论文集。89–92。[32] S.N. Kim、D. Martinez、L. Cavedon 和 L. Yencken。2011年。支持循证医学的句子自动分类。《BMC生物信息学》12，2（2011年）。[33] Grace Eunkyung Lee。2017年。用于系统综述临床文档分类的卷积神经网络研究：Sysreview 参加 CLEF eHealth 2017。载于 CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文。[34] Hang Li、Harrisen Scells 和 Guido Zuccon。2020年。用于端到端查询构建的系统综述自动化工具。载于第43届国际 ACM SIGIR 信息检索研究与发展会议论文集。2141–2144。[35] Junyi Li、Tianyi Tang、Wayne Xin Zhao、Jian-Yun Nie 和 Ji-Rong Wen。2022年。基于预训练语言模型的文本生成综述。arXiv预印本 arXiv:2201.05273（2022年）。[36] Valentin Liévin、Christoffer Egeberg Hother 和 Ole Winther。2023年。大语言模型能否对医学问题进行推理？arXiv：
2207 08143 [ ] cs Vivian Liu 和 Lydia B Chilton。2022年。文本生成图像模型提示工程设计指南。载于2022年 CHI计[37]
算机系统中的人为因素会议论文集（CHI '22）。美国计算机协会，纽约，纽约州，美国，
1 23 10 1145 3491102 3501825 –。https://doi.org/ 。/ 。Renqian Luo、Liai Sun、Yingce Xia、Tao Qin、[38]
Sheng Zhang、Hoifung Poon 和 Tie-Yan Liu。2022年。BioGPT：用于生物医学文本生成与挖掘的生成式预训练变换器。《生物信息学简报》23，6（2022年11月），bbac409。https://doi.org/10.1093/ bib/bbac409[39] Andrew MacFarlane、Tony Russell-Rose 和 Farhad Shokraneh。2022年。系统综述检索策略制定：问题、挑战与机遇。《智能系统与应用》（2022年），200091。[40] Iain J Marshall、Rachel Marshall、Byron C Wallace、Jon Brassey 和 James Thomas。2019年。快速综述得出的结果可能与系统综述不同：一项荟萃流行病学研究。《临床流行病学杂志》109（2019年），30–41。[41] Iain J Marshall、Anna Noel-Storr、Joël Kuiper、James Thomas 和 Byron C Wallace。2018年。用于识别随机对照试验的机器学习：评估与实践指南。《研究综合方法》9，4（2018年），602–614。[42] D. Martinez、S. Karimi、L. Cavedon 和 T. Baldwin。2008年。利用排序文本检索与分类促进生物医学系统综述。载于第13届澳大拉西亚文档计算研讨会论文集。[43] Adamantios Minas、Athanasios Lagopoulos 和 Grigorios Tsoumakas。2018年。亚里士多德大学在2018年 CLEF eHealth实验室实证医学技术辅助综述任务中的方法。。载于 CEUR研讨会论文集：CLEF 2018评测论坛会议与实验室工作论文。[44] M. Miwa、J. Thomas、A. O'Mara-Eves 和 S. Ananiadou。2014年。通过基于确定性的筛选减少系统综述工作量。JBI 51（2014年），242–253。[45] Rodrigo Nogueira、Wei Yang、Kyunghyun Cho 和 Jimmy Lin。2019年。使用 BERT 进行多阶段文档排序。arXiv预印本 arXiv:1910.14424（2019年）。[46] Christopher Norman、Mariska Leeflang 和 Aurélie Névéol。2018年。LIMSI@ CLEF eHealth 2018 任务2：通过叠加主动学习与静态学习开展技术辅助综述。。载于 CEUR研讨会论文集：CLEF 2018评测论坛会议与实验室工作论文。[47] Christopher Norman12、Mariska Leeflang 和 Aurélie Névéol。2017年。Limsi@ Clef Ehealth 2017 任务2：用于自动文章排序的逻辑回归。载于 CEUR研讨会论文集：CLEF 2019评测论坛会议与实验室工作论文。[48] Jo nas Oppenlaender。2022年。文本生成图像的提示修饰符分类体系。arXiv:2204.13988 [cs]

[49] Mohammadreza Pourreza 和 Faezeh Ensan。2022年。面向生物医学系统文献综述的语义驱动布尔查询形式化。《国际医学信息学杂志》（2022年），104928。[50] Animesh Prasad、Manpreet Kaur 和 Min-Yen Kan。2018年。Neural ParsCit：一种基于深度学习的参考文献字符串解析器。《数字图书馆杂志》19（2018年），323–337。[51] John Rathbone。2017年。系统综述自动化。博士学位论文。邦德大学。[52] Scott Reeves、Ivan Koppel、Hugh Barr、Della Freeth 和 Marilyn Hammick。2002年。开展系统综述的十二条建议。《医学教师》24，4（2002年）。[53] Laria Reynolds 和 Kyle McDonell。2021年。大语言模型提示编程：超越少样本范式。载于2021年 CHI计算机系统中的人为因素会议扩展摘要。ACM，横滨，日本，1–7。https://doi.org/10.1145/ 3411763.3451760[54] Raul Rodriguez-Esteban 和 Ivan Iossifov。2009年。生物医学研究中的图表挖掘。《生物信息学》25，16（2009年），2082–2084。[55] Tony Russell-Rose 和 Philip Gooch。2018年。2dSearch：一种检索策略制定的可视化方法。载于第1届实验性搜索与信息检索系统设计双年会议论文集。[56] José Antonio Salvador-Oliván、Gonzalo Marco-Cuenca 和 Rosario Arquero-Avilés。2019年。系统综述所用检索策略中的错误及其对信息检索的影响。《医学图书馆协会杂志》：JMLA 107，2（2019年4月），210–221。https://doi.org/ 10 5195 2019 567 。/jmla. 。Margaret Sampson 和 Jessie McGowan。2006年。检索策略中的错误按类型[57]和频率进行了识别。《临床流行病学杂志》59，10（2006年），1057–e1。[58] Harrisen Scells、Connor Forbes、Justin Clark、Bevan Koopman 和 Guido Zuccon。2022年。查询优化对系统综述文献检索的影响：查询日志分析。载于2022年 ACM SIGIR信息检索理论国际会议论文集。34–42。[59] Harrisen Scells 和 Guido Zuccon。2018年。为系统综述生成更好的查询。载于第41届国际 ACM SIGIR 信息检索研究与发展会议论文集（美国密歇根州安娜堡）（SIGIR '18）。10页。[60] Harrisen Scells 和 Guido Zuccon。2018年。searchrefiner：用于系统综述的查询可视化与理解工具。载于第27届 ACM信息与知识管理国际会议论文集。ACM，1939–1942。[61] Harrisen Scells、Guido Zuccon、Anthony Deacon 和 Bevan Koopman。2017年。QUT Ielab 参加 CLEF eHealth 2017技术辅助综述赛道：学习排序的初步实验。载于 CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文。[62] Harrisen Scells、Guido Zuccon 和 Bevan Koopman。2019年。系统综述文献检索的自动布尔查询优化。载于万维网大会论文集。1646–1656。[63] Harrisen Scells、Guido Zuccon 和 Bevan Koopman。2020年。系统综述自动布尔查询构建方法的对照。《信息检索杂志》（2020年），1–26。[64] Harrisen Scells、Guido Zuccon、Bevan Koopman 和 Justin Clark。2020年。系统综述文献检索的自动布尔查询构建。载于2020年万维网大会论文集。1071–1081。[65] Harrisen Scells、Guido Zuccon、Bevan Koopman 和 Justin Clark。2020年。客观导出系统综述检索策略的计算方法。载于欧洲信息检索会议。施普林格，385–398。[66] Harrisen Scells、Guido Zuccon、Mohamed A Sharaf 和 Bevan Koopman。2020年。对查询变体进行抽样以用于学习排序，从而改进系统综述中的自动布尔查询生成。载于2020年万维网大会论文集。3041–3048。[67] Michael Simon、Elke Hausner、Susan F Klaus 和 Nancy E Dunton。2010年。通过 Medline 识别护士人力配置研究：使用 PubMed 界面制定并测试基于实证导出的检索策略。《BMC医学研究方法学》10，1（2010年），76。[68] Gaurav Singh、Iain Marshall、James Thomas 和 Byron Wallace。2017年。使用深度模型识别诊断试验准确性研究文献。载于 CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文，第1866卷。[69] Jaspreet Singh 和 Lini Thomas。2017年。IIIT-H 参加 CLEF eHealth 2017 任务2：实证医学技术辅助综述。。载于 CEUR研讨会论文集：CLEF 2017评测论坛会议与实验室工作论文。[70] Claire Stansfield、James Thomas 和 Josephine Kavanagh。2013年。自动对文档进行“聚类”以支持研究范围综述：案例研究。《研究综合方法》4，3（2013年），230–241。[71] Rodney Summerscales、Shlomo Argamon、Jordan Hupert 和 Alan Schwartz。2009年。在医学摘要中识别治疗措施、组别和结局。载于第6届中西部计算语言学研讨会论文集。[72] Rodney L Summerscales、Shlomo Argamon、Shangda Bai、Jordan Hupert 和 Alan Schwartz。2011年。临床试验结果的自动摘要。载于2011年 IEEE生物信息学与生物医学国际会议论文集。372–377。[73] James Thomas 和 Angela Harden。2008年。系统综述中定性研究的主题综合方法。《BMC医学研究方法学》8，1（2008年），45。[74] Prem Timsina、Jun Liu 和 Omar El-Gayar。2016年。医学系统综述自动化的高级分析。《信息系统前沿》18，2（2016年），237–252。

[75] Mercedes Torres Torres 和 Clive E Adams。2017年。RevManHAL：迈向系统综述中的自动文本生成。《系统综述》6, 1 (2017年)。[76] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。Attention Is All You Need。神经信息处理系统进展30 (2017年)。[77] Shuai Wang、Hang Li、Harrisen Scells、Daniel Locke 和 Guido Zuccon。2021年。面向系统综述文献检索的 MeSH 主题词推荐。载于《第*25*届澳大拉西亚文档计算研讨会论文集》。1–8。[78] Shuai Wang、Hang Li 和 Guido Zuccon。2022年。MeSH 推荐器：用于系统综述布尔查询构建的 MeSH 主题词推荐库和系统。*arXiv* 预印本 *arXiv:2212.09018*（2022年）。[79] Shuai Wang、Harrisen Scells、Justin Clark、Bevan Koopman 和 Guido Zuccon。2022年。小事成就大事：用于医学系统综述文献检索的种子研究合集。载于国际 *ACM SIGIR* 信息检索研究与发展会议第*45*届会议论文集。3176–3186。[80] Shuai Wang、Harrisen Scells、Bevan Koopman 和 Guido Zuccon。2022年。医学系统综述文献检索中有效查询构建的自动 MeSH 主题词推荐。《智能系统与应用》（2022年），200141。[81] Huaying Wu、Tingting Wang、Jiayi Chen、Su Chen、Qinmin Hu 和 Liang He。2018年。华东师范大学参加 2018年电子健康任务 2：实证医学技术辅助综述。《酶学方法》的配套期刊《*Methods-a*》4, 5 (2018年)，7。[82] Zhe Yu 和 Tim Menzies。2017年。技术辅助综述的数据平衡：欠采样还是重新加权……载于*CEUR*研讨会论文集：*CLEF 2017*：评测论坛会议与实验室工作记录。