通过提示大语言模型进行查询扩展

Rolf Jagerman 谷歌研究院
jagerman@google.com

Honglei Zhuang
Google Research
hlz@google.com

Zhen Qin
谷歌研究院
zhenqin@google.com

Xuanhui Wang
Google 研究院
xuanhui@google.com

Michael Bendersky
谷歌研究
bemike@google.com

arXiv:2305.03653v1 [cs.IR] 5 May 2023

摘要

查询扩展是一种广泛使用的技术，可用于提高搜索系统的召回率。本文提出了一种利用大语言模型（LLM）的生成能力进行查询扩展的方法。与传统的查询扩展方法（如伪相关反馈（PRF））不同，后者依赖于检索一组高质量的伪相关文档来扩展查询，而我们的方法依赖于大语言模型的生成与创造能力，并利用模型内在的知识。我们研究了多种不同的提示策略，包括零样本、少样本和思维链（CoT）提示。我们发现，思维链提示对查询扩展尤为有用，因为这类提示能够引导模型逐步分解查询，并提供大量与原始查询相关的术语。在 MS-MARCO 和 BEIR 数据集上的实验结果表明，大语言模型生成的查询扩展内容可能比传统查询扩展方法更有效。

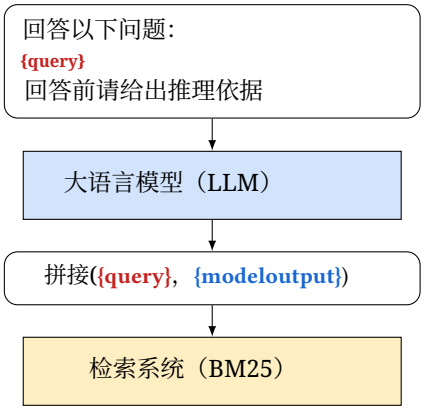

图 1：使用零样本思维链（CoT）提示词生成查询扩展词的高层概览。

1 引言

查询扩展是一种广泛使用的技术，它通过向原始查询添加额外的词语来提高搜索系统的召回率。扩展后的查询或许能够找回与原始查询没有词汇重叠的相关文档。传统的查询扩展方法通常基于伪相关反馈（PRF） [1, 20, 21, 23], ，即将原始查询检索到的文档集视为“伪相关”文档，并利用这些文档的内容提取新的查询词。然而，基于 PRF 的方法假设排名靠前的检索文档与查询相关。实际上，初始检索到的文档可能与原始查询并不完全匹配，尤其是在查询较短或含义模糊时。因此，如果初始检索文档集质量不够理想，基于 PRF 的方法就可能失效。

本文提出利用大语言模型（LLM） [3, 8, 19]辅助查询扩展。近年来，大语言模型在信息检索（IR）领域受到越来越多的关注。它们具备回答问题、生成文本等多种能力，因此是强大的工具。我们提出利用这些生成能力来生成有用的查询扩展。具体而言，我们研究如何通过提示词引导大语言模型，为原始查询生成各种替代表达和新词语。这意味着，我们不再依赖 PRF 文档或词汇知识库中的知识，而是利用大语言模型本身蕴含的知识。图 1 展示了所提出方法论的一个示例。

本研究的主要贡献如下：首先，我们设计了多种提示词来进行查询扩展（零样本、少样本和思维链），并分别采用和不采用 PRF，以研究它们的相对性能。其次，我们发现思维链提示的表现最佳，并推测这是因为思维链提示会指示模型逐步拆解答案，其中包含许多有助于查询扩展的关键词。最后，我们研究了不同模型规模下的性能，以更好地了解基于大语言模型的查询扩展方法在实际应用中的能力与局限性。

2 相关工作

查询扩展已得到广泛研究 [4, 11]。从本质上说，查询扩展通过将查询词扩展为表达相同概念或信息需求的新词，帮助检索系统提高其与语料库中文档进行词汇匹配的可能性。早期关于查询扩展的研究主要聚焦于使用词汇知识库 [2, 18, 29] 或伪相关反馈（PRF） [1, 20, 23]。基于 PRF 的方法在实践中特别有用，因为它们无需构建特定领域的知识库，且可应用于任何语料库。与查询扩展正交的是文档扩展 [10, 16, 25, 33]，它采用类似技术，但在索引阶段扩展文档词项，而不是在检索阶段扩展查询词项。

近期关于查询扩展的研究利用神经网络生成或选择扩展词，通常通过训练或微调某个模型来实现。与之不同，我们的工作利用通用 大语言模型固有的能力，无需训练或微调模型。

我们注意到，我们的工作与 [7] 和[31]近期开展的工作相似：利用大语言模型扩展查询。不过，我们的工作在几个重要方面有所不同：首先，我们研究了多种不同的提示，而 [31] 聚焦于单个少样本提示，[7] 则没有研究提示。其次，与 [31] 和 [7],不同，我们专注于生成查询扩展词，而非完整的伪文档。为此，我们展示了提示词在多种较小模型规模下的表现，从而帮助理解基于大语言模型的查询扩展方法的局限性和实际能力。最后，我们使用完全开源的模型进行实验，以促进研究的可复现性与开放性；而 [31] 仅使用一种只能通过第三方 API 访问的模型进行实验。

3 方法论

我们将查询扩展问题表述如下：给定一个查询q，我们希望生成一个扩展后的查询 q'，其中包含可能有助于检索相关文档的其他查询词。具体来说，我们研究使用大语言模型扩展查询词并生成新查询 q'。由于大语言模型的输出可能较冗长，我们将原始查询词重复 5 次，以提高其相对重要性。这与 [31] 采用的技巧相同。更正式地说：

$$q' = \text{Concat}(q, q, q, q, q, \text{LLM}(prompt_q)), \quad (1)$$

其中 Concat 是字符串拼接运算符，q 是原始查询，LLM 是大语言模型，而提示词q 是基于查询（以及可能包含的侧信息，如少样本示例或伪相关反馈文档）生成的提示词。

本文研究了八种不同的提示：

Q2D Query2Doc [31] 少样本提示词，要求模型撰写一段回答查询的文字。 **Q2D/ZS** **Q2D**的零样本版本。**Q2D/PRF** 类似于 **Q2D/ZS** 的零样本提示词，但还包含额外的上下文，即为该查询检索到的排名前三的 PRF 文档。**Q2E** 与 Query2Doc 少样本提示词类似，但示例内容是查询的扩展词，而不是文档。 **Q2E/ZS** **Q2E**的零样本版本。**Q2E/PRF** 类似于**Q2E/ZS** 的零样本提示词，但包含类似**Q2D/PRF**的额外上下文，即 PRF 文档。 **CoT** 一种零样本思维链提示词，指示模型为其回答提供推理依据。**CoT/PRF** 类似于**CoT** 的提示词，但还包含额外的上下文，即为该查询检索到的排名前三的 PRF 文档。

零样本提示（**Q2D/ZS** 和 **Q2E/ZS**）是最简单的，因为它们由一条简单的纯文本指令和输入查询组成。少样本提示（**Q2D**和 **Q2E** ）还包含若干示例，以支持上下文学习，例如，其中包含查询及其对应的扩展。思维链（**CoT**）提示通过要求模型逐步拆解回答来撰写指令，以获得更详细的输出。最后，伪相关反馈（·/**PRF**）提示变体将排名前 3 的检索文档用作额外上下文，供模型使用。实验中使用的确切提示词见附录A。

4 实验

为验证基于大语言模型的查询扩展的有效性，我们在两项检索任务上开展实验：MS-MARCO [15] 段落检索和 BEIR [27] 25 [21, 。检索系统采用 22] [17][1] 25 Terrier 实现的 BM。我们使用 Terrier 提供的默认 BM 参数（$b = 0.75$、 $k_1 = 1.2$、$k_3 = 8.0$）。

4.1 基线

为分析基于大语言模型的查询扩展方法，我们将其与几种经典的基于 PRF 的查询扩展方法进行比较 [1]:

- Bo1：玻色-爱因斯坦 1 加权
- Bo2：玻色-爱因斯坦 2 加权
- KL：Kullback–Leibler 加权

这些方法的实现由 Terrier 提供。在所有情况下，我们都使用 Terrier 的默认查询扩展设置：3 篇 PRF 文档和 10 个扩展词。

此外，我们将 Query2Doc [31] 中的提示作为基线。不过，我们不与其确切设置进行比较，因为他们使用的模型规模远大于本文研究的模型。本文的比较侧重于提示，而非不同模型（其中可能有规模大得多的模型）所生成的具体数值。此外，对于感受野较小的模型（具体来说是 Flan-T5 模型），我们仅使用 3-shot Q2D 提示，而非标准的 4-shot 提示，以防提示被截断。

4.2 语言模型

我们在两类不同规模的模型上比较这些提示：Flan-T5 [6, 19]和 Flan-UL2 [26],:

- Flan-T5-Small（6000 万个参数）
- Flan-T5-Base（220M 个参数）
- Flan-T5-Large（770M 个参数）
- Flan-T5-XL（3B 个参数）
- Flan-T5-XXL（11B 个参数）
- Flan-UL2（20B 个参数）

我们选择使用 T5 [19] 和 UL2 [26] 模型的 Flan [6, 32] 版本，因为这些模型经过微调，能够遵循指令；在使用基于提示的方法时，这一点至关重要。此外，这些模型均以开源形式提供[2]。

4.3 评估指标

由于我们关注查询扩展，而查询扩展主要致力于提升第一阶段检索的召回率，因此我们使用 Recall@1K 作为核心评估指标。我们还使用 MRR@10 [30] 和 NDCG@10 [14] 报告前列加权排序指标，以更好地了解模型如何改变排名靠前的检索结果。所有结果均通过配对 t-test 进行显著性检验，并将 $p < 0.01$ 作为结果显著的标准。

[1]http://terrier.org/ [2]模型可在 https://huggingface.co/docs/transformers/model_doc/flan-t5 和 https://huggingface.co/google/flan-ul2 上获取

5 结果

5.1 MS-MARCO 段落排序

表 1 展示了 MS-MARCO 段落排序任务上的结果。经典的查询扩展基线方法（Bo1、Bo2 和 KL）相较于标准 BM25 检索，已经在 Recall@1K 方面带来了有益的提升。与 [12], 的结果一致，我们观察到，召回率的提升是以牺牲 MRR@10 和 NDCG@10 等前列加权排序指标为代价的。

接下来，我们看到，基于大语言模型的查询扩展效果在很大程度上取决于所使用的提示类型。与 [31],Query2Doc 提示的研究结果类似，**Q2D** 能够在 Recall@1K 方面相较于经典方法实现显著提升。有趣的是，Query2Doc 不仅能够提高召回率，还能改善 MRR@10 和 NDCG@10 等前列加权排序指标，从而在各项指标上都带来良好的提升。这与经典查询扩展方法形成对比，后者通常会牺牲前列加权排序指标，以提升召回率。

最后，最佳性能由 **思维链** （以及相应的 PRF 增强提示 **CoT/PRF**）取得。该提示会指导模型生成详细的解释，将答案分解为多个步骤。我们推测，这种详尽性可能会产生许多对查询扩展有用的潜在关键词。最后，我们发现，在提示中加入 PRF 文档能够显著改善不同模型和提示下 MRR@10 和 NDCG@10 等前列加权排序指标。一种可能的解释是，大语言模型能够有效提炼 PRF 文档中的信息；这些文档可能已经包含相关段落，而模型则通过关注最有希望的关键词并在输出中使用它们来完成这一过程。附录 B 提供了提示输出的一个更具体示例。

5.2 BEIR

BEIR 数据集包含来自多个领域的许多不同的零样本信息检索任务。我们在表 2 中比较了不同提示在 BEIR 数据集上的表现。首先值得注意的是，传统的基于 PRF 的查询扩展基线仍然表现出色，尤其是在 trec-covid、scidocs 和 touche2020 等特定领域的数据集上。这些数据集大多具有学术和科学性质，在这些情况下，PRF 文档可能会提供有用的查询词。相比之下，通用大语言模型可能缺乏足够的领域知识，因而无法有效处理这些数据集。其次，我们注意到，问答类数据集（fiqa、hotpotqa、msmarco 和 nq）似乎最能从基于大语言模型的查询扩展方法中获益。这可能是因为语言模型会生成与查询相关的答案，从而更有效地检索相关段落。在所有数据集上，**Q2D/PRF** 提示的平均 Recall@1K 最高，**思维链** 提示则紧随其后。

5.3 模型规模的影响

为了了解基于大语言模型的查询扩展器在实际应用中的能力与局限性，我们在图 2 中比较了不同规模的模型。模型规模从 6000 万参数（Flan-T5-small）到 110 亿参数（Flan-T5-XXL）不等，我们还尝试了一个拥有 200 亿参数的模型（Flan-UL2），但需注意，该模型采用了不同的预训练目标。总体而言，我们观察到了预期趋势：规模更大的模型往往会

表 1：基于大语言模型的查询扩展在 MS-MARCO 段落排序开发集上的结果。▲ 表示相较于 Q2D Flan-UL2 方法有统计显著性提升（配对 t检验，$p < 0.01$）。每项指标的最佳结果以粗体标出。

	Recall@1K	MRR@10	NDCG@10
BM25	87.82	18.77	23.44
BM25 + Bo1	88.68	17.75	22.48
BM25 + Bo2	88.32	17.58	22.30
BM25 + KL	88.62	17.71	22.44
Flan-T5-XXL (11B)			
Q2D	88.76	19.07	23.76
Q2D/ZS	88.88	18.55	23.13
Q2D/PRF	89.31	22.13▲	26.43▲
Q2E	87.74	18.74	23.37
Q2E/ZS	87.93	18.79	23.45
Q2E/PRF	88.20	19.20	23.83
CoT	89.86	19.16	23.82
CoT/PRF	89.02	22.08▲	26.32▲
Flan-UL2 (20B)			
Q2D	89.87	19.22	23.96
Q2D/ZS	86.60	15.56	19.54
Q2D/PRF	89.28	21.42▲	25.82▲
Q2E	88.04	18.84	23.52
Q2E/ZS	88.11	18.87	23.56
Q2E/PRF	88.43	19.24	23.90
CoT	**90.61**▲	20.05▲	24.85▲
思维链/PRF	89.30	**22.62**▲	**26.89**▲

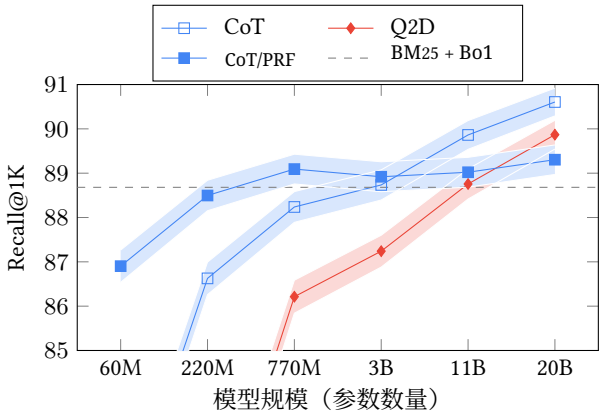

图 2：不同模型规模在 MS-MARCO 段落排序开发集上的表现。阴影区域表示 99% 置信区间。

表现更好。**Q2D** 方法至少需要一个 11B 参数的模型，才能达到 BM25+Bo1 基线的水平。相比之下，**思维链** 方法只需要一个 3B 参数的模型即可达到同等水平。此外，在**思维链** 提示词中加入伪相关反馈文档，似乎有助于稳定较小模型规模下的表现，但会抑制较大模型规模下的表现。对此现象的一种可能解释是，伪相关反馈文档会降低

表 2：使用 Flan-UL2 在 BEIR 上采用各种提示时的 Recall@1K。▲表示相较于最佳经典查询扩展方法，在统计上具有显著性（配对 t-检验，$p < 0.01$）的提升。每个数据集的最佳结果以粗体突出显示。

		经典查询扩展			基于大语言模型的查询扩展							
数据集	BM25	Bo1	Bo2	KL	Q2D	Q2D/ZS	Q2D/PRF	Q2E	Q2E/ZS	Q2E/PRF	CoT	CoT/PRF
arguana	98.93	**99.00**	**99.00**	**99.00**	98.86	98.93	98.93	98.93	98.93	98.93	98.93	98.86
climate-fever	46.60	45.69	45.38	45.65	47.62	47.66	**47.94**	46.08	46.44	46.44	47.42	46.81
cqadupstack	65.55	**66.82**	66.57	66.70	65.51	64.19	65.01	65.69	65.71	65.90	66.39	66.12
DBpedia	63.72	64.77	64.55	64.60	**65.89**	65.47	65.78	63.55	63.92	63.93	65.77	65.06
发烧	75.73	76.28	75.83	76.32	**79.06**▲	78.87▲	77.29	75.78	75.79	76.27	78.21▲	77.25
fiqa	77.42	79.18	79.06	78.84	78.34	78.26	78.69	77.33	77.31	77.68	**80.08**	79.03
HotpotQA	85.78	84.84	81.71	84.65	86.90▲	85.71	87.58▲	85.60	85.54	87.25▲	87.54▲	**88.79**▲
MS MARCO	73.61	75.08	75.14	74.66	76.77	75.73	78.75	73.87	73.79	74.14	**79.58**	78.36
NFCorpus	38.70	57.30	57.67	56.46	55.34	**59.81**	59.68	43.38	44.12	47.06	52.63	53.32
nq	78.96	81.09	80.64	80.82	85.18▲	84.71▲	83.53▲	79.30	79.11	80.35	**85.46**▲	83.11▲
Quora	99.26	99.20	99.12	99.20	99.00	98.84	98.92	99.25	**99.29**	99.26	99.17	99.21
SciDocs	57.46	59.78	**61.03**	59.86	59.09	59.78	60.10	57.88	57.70	58.32	58.51	59.69
SciFact	97.17	**97.57**	**97.57**	**97.57**	**97.57**	**97.57**	**97.57**	97.17	97.17	97.17	**97.57**	97.17
Touché2020	84.96	85.94	**86.38**	86.01	83.61	83.44	84.54	85.21	85.02	86.04	85.51	84.58
TREC-COVID	42.58	45.21	**45.58**	45.39	43.52	38.05	44.17	43.16	43.12	43.85	43.43	44.02
平均值	72.43	74.52	74.35	74.38	74.82	74.47	**75.23**	72.81	72.86	73.50	75.08	74.76

模型的创造力，因为模型可能会过度关注所提供的文档。虽然这有助于防止较小模型规模下的模型出错，但也会抑制我们希望在较大模型规模下利用的创造能力。在参数规模为 770M 的模型上，**CoT/PRF**提示的表现优于其他提示，因此它是实际搜索场景中值得考虑的候选方案；在这类场景中，可能无法部署更大的模型。总体而言，显然，大模型能够带来显著提升，但这可能会限制大语言模型查询扩展方法的实际应用。已有研究表明，蒸馏是将大模型能力迁移到小模型的有效方法。我们将查询扩展模型的蒸馏研究留作未来工作。

6 局限性与未来工作

我们的工作存在一些局限性：首先，我们仅研究稀疏检索（BM25），这是查询扩展发挥重要作用的场景。稠密检索系统（例如双编码器）不太容易受到词汇鸿沟的影响，因此从查询扩展中获益的可能性也较低。Wang等人 [31] 已经更详细地研究了这一设置，而我们将针对稠密检索设置分析我们的提示词留作未来工作。其次，我们的工作聚焦于Flan [32] 指令微调语言模型。我们选择这些模型，是因为它们能够遵循指令，而且这些模型是开源的。我们的工作可以自然地扩展到其他语言模型 [3, 5, 9, 28]，而我们将对此类模型的研究留作未来研究主题。第三，我们研究了特定的提示词模板（见附录A），而构造不同提示词可能还有其他方式。最后，大语言模型的计算成本可能过高，使得在实践中部署基于大语言模型的查询扩展变得难以实现。或许可以将大型模型的输出蒸馏到一个更小、可服务的模型中。如何将基于大语言模型的查询扩展投入生产仍是一个开放问题。

7 结论

本文研究基于大语言模型的查询扩展。与传统的基于 PRF 的查询扩展不同，大语言模型不受初始检索文档集合的限制，因此能够生成传统方法未能涵盖的扩展词项。我们提出的方法十分简单：向大语言模型提供一个提示词和一个查询，然后利用模型的输出，通过添加有助于文档检索的新词项来扩展原始查询。

我们的结果表明，思维链提示特别适用于查询扩展，因为它们会引导模型生成详尽的解释，从而涵盖多种新关键词。此外，我们的结果显示，在各种提示中加入 PRF 文档可以提升检索阶段的前列加权排序指标表现，并且在使用较小模型规模时更加稳健，这有助于大语言模型查询扩展的实际部署。

正如本文所示，查询扩展等信息检索任务可以从大语言模型中受益。随着大语言模型能力不断提升，令人期待的是，它们的能力将应用于各种信息检索任务。此外，随着大语言模型日益普及，将其用作信息检索系统的核心组成部分也会变得更加容易，这让此类系统的从业者和研究人员都感到振奋。

参考文献

[1] Gianni Amati 和 Cornelis Joost Van Rijsbergen。2002。基于随机性偏离度测量的信息检索概率模型。ACM信息系统汇刊（TOIS）20，4（2002），357–389。[2] Jagdev Bhogal、Andrew MacFarlane 和 Peter Smith。2007。基于本体的查询扩展综述。信息处理与管理 43，4（2007），866–886。[3] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020。语言模型是少样本学习者。神经信息处理系统进展 33（2020），1877–1901。[4] Claudio Carpineto 和 Giovanni Romano。2012。信息检索中自动查询扩展综述。ACM计算调查（CSUR）44，1（2012），1–50。[5] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Se- bastian Gehrmann 等。2022。PaLM：利用路径扩展语言建模规模。arXiv预印本 arXiv:2204.02311（2022）。[6] Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Eric Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。2022。指令微调语言模型的规模扩展。arXiv预印本 arXiv:2210.11416（2022）。[7] Vincent Claveau。2021。用于信息检索中查询扩展的神经文本生成。载于IEEE/WIC/ACM国际网络智能与智能代理技术会议。202–209。[8] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018。BERT：用于语言理解的深度双向变换器预训练。arXiv预印本 arXiv:1810.04805（2018）。[9] Nan Du、Yanping Huang、Andrew M Dai、Simon Tong、Dmitry Lepikhin、Yuanzhong Xu、Maxim Krikun、Yanqi Zhou、Adams Wei Yu、Orhan Firat 等。2022。GLaM：利用混合专家实现语言模型的高效规模扩展。载于国际机器学习会议。PMLR，5547–5569。[10] Miles Efron、Peter Organisciak 和 Katrina Fenlon。2012。通过文档扩展改进短文本检索。载于第35届ACM国际信息检索研究与开发会议论文集。911–920。[11] Efthimis N Efthimiadis。1996。查询扩展。信息科学与技术年度评论（ARIST）31（1996），121–87。[12] D Harman。 [无日期。]相关性反馈及其他查询修改技术。信息检索：数据结构与算法（[无日期]），241–263。[13] Ayyoob Imani、Amir Vakili、Ali Montazer 和 Azadeh Shakery。2019。利用词嵌入进行查询扩展的深度神经网络。载于《信息检索进展：第41届欧洲信息检索研究会议，ECIR 2019》，德国科隆，2019年4月14日至18日，论文集，第二部分41。Springer，203–210。[14] Kalervo Järvelin 和 Jaana Kekäläinen。2002。基于累积增益的信息检索技术评估。ACM信息系统汇刊（TOIS）20，4（2002），422–446。[15] Tri Nguyen、Mir Rosenberg、Xia Song、Jianfeng Gao、Saurabh Tiwary、Rangan Majumder 和 Li Deng。2016。MS MARCO：人工生成的机器阅读理解数据集。choice 2640（2016），660。[16] Rodrigo Nogueira、Wei Yang、Jimmy Lin 和 Kyunghyun Cho。2019。通过查询预测进行文档扩展。arXiv预印本 arXiv:1904.08375（2019）。[17] Iadh Ounis、Gianni Amati、Vassilis Plachouras、Ben He、Craig Macdonald 和 Douglas Johnson。2005。Terrier信息检索平台。载于《信息检索进展：第27届欧洲信息检索研究会议，ECIR 2005》，西班牙圣地亚哥-德孔波斯特拉，2005年3月21日至23日，论文集27。Springer，517–519。[18] Yonggang Qiu 和 Hans-Peter Frei。1993。基于概念的查询扩展。载于第16届ACM国际信息检索研究与开发年度会议论文集。160–169。[19] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。2020。探索统一文本到文本变换器的迁移学习极限。机器学习研究杂志 21，1（2020），5485–5551。[20] Stephen E Robertson。1990。关于查询扩展中的词项选择。文献学杂志 46，4（1990），359–364。[21] Stephen E Robertson 和 K Sparck Jones。1976。搜索词的相关性加权。美国信息科学学会杂志 27，3（1976），129–146。[22] Stephen E Robertson、Steve Walker、Susan Jones、Micheline M Hancock-Beaulieu、Mike Gatford 等。1995。TREC-3中的Okapi。NIST特别出版物 SP 109（1995），109。[23] Joseph John Rocchio Jr. 1971。信息检索中的相关性反馈。《SMART检索系统：自动文档处理实验》（1971）。[24] Dwaipayan Roy、Debjyoti Paul、Mandar Mitra 和 Utpal Garain。2016。利用词嵌入实现自动查询扩展。arXiv预印本 arXiv:1606.07608（2016）。[25] Tao Tao、Xuanhui Wang、Qiaozhu Mei 和 ChengXiang Zhai。2006。结合文档扩展的基于语言模型的信息检索。载于……

NAACL 人类语言技术会议，主会议。407–414。[26] Yi Tay、Mostafa Dehghani、Vinh Q Tran、Xavier Garcia、Dara Bahri、Tal Schus- ter、Huaixiu Steven Zheng、Neil Houlsby 和 Donald Metzler。2022。统一语言学习范式。*arXiv* 预印本 *arXiv:2205.05131*（2022）。[27] Nandan Thakur、Nils Reimers、Andreas Rücklé、Abhishek Srivastava 和 Iryna Gurevych。2021。BEIR：用于零样本评估信息检索模型的异构基准。在神经信息处理系统第三十五届会议数据集与基准赛道（第二轮）。https://openreview. net/forum?id=wCu6T5xFjeJ [28] Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kul- shreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du 等。2022。Lamda：用于对话应用的语言模型。*arXiv* 预印本 *arXiv:2201.08239*（2022）。[29] Ellen M Voorhees。1994。使用词汇-语义关系进行查询扩展。在*SIGIR'94*：第十七届 *ACM-SIGIR* 信息检索研究与开发年度国际会议论文集，由都柏林城市大学组织。Springer，61–69。[30] Ellen M Voorhees 等。1999。TREC-8 问答赛道报告。在*Trec*，第 99 卷。77–82。[31] Liang Wang、Nan Yang 和 Furu Wei。2023。Query2Doc：使用大语言模型进行查询扩展。*arXiv* 预印本 *arXiv:2303.07678*（2023）。[32] Jason Wei、Maarten Bosma、Vincent Y Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M Dai 和 Quoc V Le。2021。微调语言模型是零样本学习器。*arXiv* 预印本 *arXiv:2109.01652*（2021）。[33] Zhi Zheng、Kai Hui、Ben He、Xianpei Han、Le Sun 和 Andrew Yates。2020。BERT-QE：用于文档重排序的上下文化查询扩展。*arXiv* 预印本 *arXiv:2009.07258*（2020）。[34] Zhi Zheng、Kai Hui、Ben He、Xianpei Han、Le Sun 和 Andrew Yates。2021。通过无监督片段选择实现上下文化查询扩展，用于文本检索。信息处理与管理58，5（2021），102672。

A 提示

表 3 列出了本文尝试过的所有提示。在每个提示中，{query} 表示我们希望为其生成查询扩展的查询。我们用{query 1}、. . .、{query 4} 表示来自 MS-MARCO 训练集的示例查询。同样，{doc 1}、. . .、{doc 4} 表示与这些示例查询对应的相关段落；而{expansion 1}、. . .、{expansion4}表示使用 Terrier KL 方法从这些相关段落生成的对应扩展（最多 20 个词）。最后，我们用{PRF doc 1}、. . .、{PRF doc3} 表示使用原始查询检索得到的排名前 3 的检索文档，这些文档充当伪相关反馈文档。对于思维链提示，我们注意到模型往往会在末尾输出“The final answer:”或“So the final answer is:”，因此在将模型输出与查询拼接之前，我们会过滤掉这两个句子。

表 3：各种查询扩展提示词。

ID	提示词
Q2D [31]	撰写一段能够回答给定查询的段落： 查询：{query1} 段落：{doc 1} 查询：{query2} 段落：{doc 2} 查询：{query3} 段落：{doc 3} 查询：{query4} 段落：{doc 4} 查询：{query} 段落：
Q2D/ZS	撰写一段能够回答以下查询的文字：{query}
Q2D/PRF	请根据以下内容撰写一段回答给定查询的文字 上下文： 上下文：{PRF doc 1} {PRF doc 2} {PRF doc 3} 查询：{query} 段落：
Q2E	为给定的查询撰写一组关键词： 查询：{query1} 关键词：{expansion 1} 查询：{query2} 关键词：{expansion 2} 查询：{query3} 关键词：{expansion 3} 查询：{query4} 关键词：{expansion 4} 查询：{query} 关键词：
Q2E/ZS	<escape>为以下查询编写一个关键词列表：{query}</escape>
Q2E/PRF	根据给定的查询编写关键词列表。 上下文： 上下文：{PRF doc 1} {PRF doc 2} {PRF doc 3} 查询：{query} 关键词：
CoT	回答以下查询： {查询} 请在回答前给出推理依据。
CoT/PRF	根据上下文回答以下查询： 上下文：{PRF 文档 1} {PRF 文档 2} {PRF 文档 3} 查询：{query} 请在回答前给出推理依据。

B 示例输出

表 4 显示了使用 Flan-T5-Large（770M）模型规模和 Flan-UL2（20B）模型规模进行查询扩展的结果。首先请注意，在较小的模型规模下，思维链和 Q2D 提示词未能给出正确答案，这会损害检索性能。思维链/伪相关反馈提示以其 PRF 文档为依据，因此避免了这一问题，并正确生成答案“塔塔汽车”，有助于检索相关段落。在较大的模型规模（Flan-UL2）下，Q2D、思维链和思维链/伪相关反馈提示都能给出正确答案“塔塔汽车”。不过，思维链提示词提供了最详尽的答案解释，与相关段落有许多术语重合，因此提升了整体检索性能。

表4：使用 Flan-UL2 进行查询扩展的示例。绿色高亮显示正确答案，蓝色表示特定术语重叠，表明这些查询扩展可能有用。

	查询 段落	谁拥有捷豹汽车？(ID=1045405) 捷豹是捷豹 路虎 英 其总部位于惠特利， 英格兰 文垂，由印度企业集团所拥有—— 业塔塔 以来一直拥有。
Flan-T5-Large (770M)	Q2D	捷豹汽车是一家英国豪华汽车制造商—— 总部位于英国。
	CoT	捷豹路 所有者。答案是 答案：捷豹路虎。
	CoT/PRF	相关信息是：捷豹归.....所有 由印度汽车制造商 塔塔 Mo s有限公司。因此，最终答案是印度人 汽车制造商塔塔汽车有限
Flan-UL2 (20B)	Q2D	捷豹是塔塔汽车旗下的一个 是一家公司 由塔塔家族所有。
	CoT	捷豹路 国的 汽 威廉·里昂斯创立 于1931年。其总部位于惠特 特里，英国，并且是 富时250指数的成分股。该公司是一家全资 旗下子公司 印度塔塔
	CoT/PRF	最终答案是塔塔汽 捷豹归印度汽车制- 造商塔塔 司所有。最终答案： 塔塔 公司