Modeling and Simulation 建模与仿真, 2025, 14(4), 901-913
Published Online April 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.144340

Hans汉斯

基于检索增强生成与软提示优化的大模型开放域问答方法

刘浩然

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年3月25日；录用日期：2025年4月18日；发布日期：2025年4月28日

摘　要

针对大语言模型(LLM)在开放域问答任务中长尾知识处理能力不足的问题，本文提出了一种融合检索增强生成(RAG)与软提示优化的新型框架SOFTRAG，旨在提升模型对低频知识的利用效率并缓解传统方法的局限性。研究结合检索增强生成(RAG)与软提示优化技术，并引入基于Perceiver的软提示适配器用于提取关键信息，同时采用LoRAMoE方法实现参数高效微调。在PopQA、TriviaQA、PubHealth和ASQA等数据集上，SOFTRAG框架在准确率、推理精度及泛化能力上均显著超越无检索基线和传统RAG方法。消融实验进一步验证了软提示、检索模块和微调技术对性能提升的关键作用。本研究方法有效平衡了性能与资源开销，显著改善了大模型在处理长尾知识任务中的表现，为开放域问答提供了新的优化思路。

关键词

长尾知识，检索增强生成，软提示，大语言模型

LLMs Open-Domain Question Answering Method Based on Retrieval-Augmented Generation and Soft Prompt Optimization

Haoran Liu

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Mar. 25th, 2025; accepted: Apr. 18th, 2025; published: Apr. 28th, 2025

Abstract

To address the limitations of large language models (LLMs) in handling long-tail knowledge for

文章引用: 刘浩然. 基于检索增强生成与软提示优化的大模型开放域问答方法[J]. 建模与仿真, 2025, 14(4): 901-913.
DOI: 10.12677/mos.2025.144340

open-domain question answering tasks, this paper proposes SOFTRAG, a novel framework that integrates Retrieval-Augmented Generation (RAG) with soft prompt optimization. The framework aims to enhance the utilization efficiency of low-frequency knowledge and mitigate the constraints of traditional approaches. The study combines RAG with soft prompt optimization techniques, introducing a Perceiver-based soft prompt adapter for extracting critical information and employing the LoRAMoE method for parameter-efficient fine-tuning. Evaluated on datasets including PopQA, TriviaQA, PubHealth, and ASQA, the SOFTRAG framework demonstrates significant improvements in accuracy, reasoning precision, and generalization capabilities compared to retrieval-free baselines and conventional RAG methods. Ablation experiments further validate the critical contributions of soft prompting, retrieval modules, and fine-tuning techniques to performance enhancement. This approach effectively balances performance with computational resource requirements, substantially improving LLMs' performance on long-tail knowledge tasks and offering new optimization insights for open-domain question answering.

Keywords

Long-Tail Knowledge, Retrieval-Augmented Generation, Soft Prompt, LLMs

1. 引言

近年来，大语言模型(LLM)在多种任务中表现出色，通过在大规模数据集上的预训练，这些模型已积累了丰富的事实知识，在许多知识密集型任务中展现了出色的细粒度记忆和推理能力[1] [2]。在开放域问答(Open-Domain QA)领域，LLM 的应用不仅推动了技术的快速发展，也为知识探测提供了新的思路和方法[3]。与封闭域问答(Closed-Domain QA)不同，开放域问答不依赖于特定领域的知识库或预定义的问答对，而是需要模型能够处理各种类型的问题，并从广泛的文本资源中找到答案。

长尾数据分布是开放域问答任务面临的一项固有挑战，即仅少量头部知识有大量数据实例，而大多数长尾知识只有极少数据实例的不平衡情况。LLM 往往只能在包含大量头部知识上取得良好表现，对于长尾知识，模型难以对其进行有效记忆和推理[4]。长尾知识总量巨大，在特定任务和实际应用中发挥着重要的作用，例如，在自然语言处理中，一些罕见的命名实体、专业术语或小众领域的知识，虽然在整体数据中出现频率较低，但在特定场景下却极为关键[5]。

检索增强生成(RAG)为解决模型在长尾知识上表现不佳的问题提供了一种高效的解决方案。RAG 通过动态检索和整合外部知识来增强模型的性能。这种方法不仅提高了模型对长尾知识的处理能力，还显著提升了模型在开放域问答任务中的总体性能[6]。例如，在处理一些涉及罕见命名实体或专业术语的问题时，RAG 方法能够通过检索到的相关文档，为模型提供必要的背景知识，从而生成更准确的答案。大多数 RAG 方法在知识探测任务中表现优异，尤其在长尾知识关系的推理上展现了优势，但检索到的文档片段可能包含噪声或与查询不完全相关的信息[7]，这可能会影响模型的生成质量，如何处理这些噪音成为了一个主要的问题。SELFRAG [8]提出了一种“自反思”的策略，使模型能够根据任务需求动态检索文档，并在生成过程中评估检索到的文档相关性。本文采用类似的方式，利用模型自身的推理能力对检索的文档进行去噪处理，提升 RAG 模型在处理噪声输入时的鲁棒性，并提高其在知识密集型任务中的生成准确性。

提示学习通过任务适配的提示设计显著提升了大语言模型的实用性和灵活性[9]，但离散型提示的效

能受限于长文本处理场景中的信息衰减现象。研究表明，当 LLM 面对较长的提示时，模型倾向于遗忘部分信息，尤其是那些位于提示中间部分的信息[10]。这种遗忘现象并非简单的信息丢失，而是一种选择性关注机制，模型更倾向于关注提示的开头和结尾部分，而对中间部分的信息处理相对不足。这种现象在处理长文本时尤为明显，例如在长篇问答、文本生成或文档摘要等任务中，模型可能会遗漏关键信息，导致生成的内容不完整或不准确。此外，大规模的数据虽然可以提供丰富的上下文信息，但模型在处理这些信息时往往会面临计算资源的限制和信息处理的瓶颈。因此，简单地输入大规模自然语言文本并非一种高效且实用的策略。在此背景下，软提示技术作为提示学习的重要演进方向，通过“可学习的向量”替代传统文本指令，以任务特定的方式优化，减少模型对不相关信息的关注，提高知识利用效率[11]。

针对上述问题，本文提出了 SOFTRAG 框架。该框架利用模型自身对检索到的文档进行显式去噪处理，并利用 Perceiver 适配器将去噪后的文本嵌入向量转化为可训练的软提示(soft prompt)，以此提取出文本中的关键信息。最终，软提示与任务相关提示拼接后输入冻结的大语言模型，并结合 LoRAMoE [12]方法进行参数高效微调，从而增强模型对长尾知识的学习能力，同时提高任务适配性。参数高效微调(PEFT) [13]方法通过仅调整模型的部分参数，显著降低了计算和存储成本，同时保留了大语言模型的通用知识，并有效缓解知识灾难性遗忘问题，提升模型对头部知识的泛化能力。

本文的主要贡献如下：

- 本文提出 SOFTRAG 框架，将 RAG 与软提示优化相结合，通过参数高效微调使模型在强化长尾知识学习的同时，增强任务适配能力。
- 本文引入软提示适配器，通过注意力机制提取长提示中的关键信息，逐层过滤冗余信息，获取易于处理且信息密集的“软提示”。
- 本文在多个基准数据集上对框架性能进行全面评估，通过多种实验设置验证其有效性，并对模型表现进行深入分析。

2. RAG 与软提示优化的模型框架

本节介绍本文提出的新框架。该框架将模型检索到的文本信息转化为软提示向量，在不牺牲模型的上下文学习能力的同时，对模型进行高效微调，确保模型能充分学习长尾知识的同时又不会遗忘原本的头部知识。

2.1. 问题定义与概述

对于开放域问答任务，给定一个问题对 $T=(q,a)$，通过检索系统 Retrieve 从 Wikipedia 语料库中检索出 K 个相关文档 $D=\{d_1,d_2,\cdots,d_k\}$。由大模型 M 对检索到的文档 D 和问题对 T 进行去噪处理，生成对应的推理 $R=\{r_1,r_2,\cdots,r_k\}$。然后使用软提示适配器将推理 R 压缩为软提示 $P=\{p_1,p_2,\cdots,p_k\}$。

模型的目标是根据软提示 P 和自身的参数知识预测给定问题 q 的正确答案 a，记作 $p_\theta(a|q,P)$。此任务的主要挑战在于处理长尾知识，即需要模型有效整合流行实体和不常见实体的信息，从而提供准确答案。同时，为避免 RAG 方法中噪声的干扰，选用现有的检索器，将所有检索到的文档作为问题的输入，未进行过滤或重新排序，以最大限度发挥模型的自适应检索能力。

2.2. 框架描述

图 1 展示了整个框架的结构，由四个部分组成：1) 文档检索；2) 推理生成；3) 软提示适配；4) 模型微调。该框架首先利用文档检索引入外部知识，再通过推理生成模块对检索结果进行显式去噪，然后结合软提示适配器对推理结果进行压缩，最后将软提示和任务提示结合输入大语言模型进行参数高效微调，显著增强大模型在长尾知识任务上的表现。

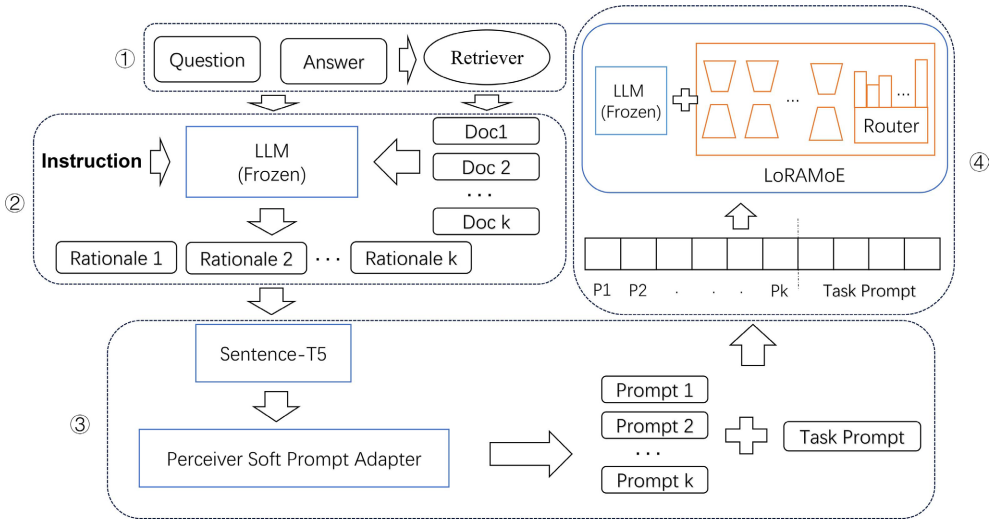

Figure 1. SOFTRAG framework: 1) Document retrieval; 2) Reasoning generation; 3) Soft prompt adaptation; 4) Model fine-tuning
图 1. SOFTRAG 框架：1) 文档检索；2) 推理生成；3) 软提示适配；4) 模型微调

2.2.1. 文档检索

在文档检索阶段，使用检索器 Retrieve 从 Wikipedia 所有文档 $W=\{w_1,w_2,\cdots,w_k\}$ 中获取与问题最相关的 K 个文档 $D=\{d_1,d_2,\cdots,d_k\}$。具体而言，Retrieve 首先通过文档编码器 $R_{\text{doc}}(\cdot)$ 和查询编码器 $R_{\text{query}}(\cdot)$ 将文档和问题分别进行编码，两个编码器都是 6 层的 Transformer，但参数不共享，并取平均池化作为最终表示；再计算出问题与文档间的余弦相似度，并从所有文档中选出 Top-K 个与问题相似度最高的文档。检索过程定义为：

$$D=\text{Retrieve}(T,W,K)=\left\{d_{(j)}\middle|\, j=\underset{i\in\{1,\ldots,N\}}{TopK}\left(sim\left(R_{doc}(w_i),R_{query}(T)\right)\right)\right\} \tag{1}$$

其中 T 是输入的问题对，$sim(\cdot)$ 表示余弦相似度。

2.2.2. 推理生成

原理生成

Input: Read the following documents relevant to the given question: $\{q\}$
Document [1] (Title:…):
{contents of d_1}
…
Document [i] (Title:…):
{contents of d_i}
…

Task-specific Instruction:
Remove irrelevant details and only retain paragraphs that directly describe the given question: $\{q\}$

Output:
{rationale r_1}
…
{rationale r_i}
…

Figure 2. Example of reasoning generation prompt
图 2. 推理生成提示样例

本研究利用大语言模型的“自反馈”能力，将问题 q 与检索到的文档集合 $D=\{d_1,d_2,\cdots,d_k\}$ 共同输入大语言模型 M，并附加任务特定指示 I，让模型做出反馈，对文档进行去噪和总结。如图 2 所示，对于每个检索到的文档 $d_i \in D$，通过大语言模型 M 的反馈，生成对应的推理 $R=\{r_1,r_2,\cdots,r_k\}$。推理生成过程为：

$$r_i = M(d_i, T, I), \forall i \in \{1,2,\cdots,k\} \tag{2}$$

I 为任务特定的指示，用于提供显式去噪信号，帮助模型聚焦于相关信息并忽略噪声。

2.2.3. 软提示适配

本研究使用 Sentence-T5 [14]作为文本编码器。在处理长文本时，Sentence-T5 可通过分块编码再融合的策略(分块输入后平均池化)，在计算资源有限时仍能保留关键信息，避免直接截断导致的语义损失。由大语言模型生成的推理 r_i 输入到 Sentence-T5 编码器(ST5)中，生成高质量句子嵌入。再将得到的嵌入输入到基于 Perceiver 模块[15]的软提示适配器 A 中，压缩为一组软提示向量 P。软提示生成过程为：

$$P = \sum_{i=1}^{k} A(\text{Sentence} - T5(r_i)) \tag{3}$$

软提示向量 P 捕捉了推理 r_i 中的关键信息，同时降低输入维度，便于与任务提示结合，从而提高模型效率。软提示的数量与检索到的 Top-K 文档数量一致。

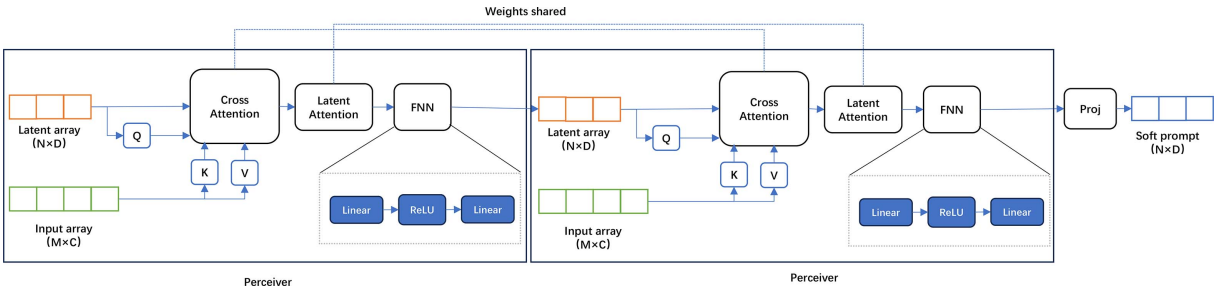

Figure 3. Perceiver-Adapter architecture
图 3. Perceiver-Adapter 架构

为了进一步提取深层次特征，本研究采用基于 Perceiver 模块的软提示适配器处理高维数据。Perceiver 通过交叉注意力机制，强制模型在压缩过程中动态关注输入文本的重要部分，且多轮潜在向量迭代处理逐步剔除冗余信息，保留与任务相关的关键特征，最终将 r_i 对应的长向量压缩为易于处理的紧凑且信息密集的“推理软提示”。这些软提示保留了输入数据的关键信息，同时显著降低了计算复杂度，非常适合用于对齐和缩减输入序列长度，避免了大模型因过长的输入而把注意力分散到无关紧要的信息。

Perceiver-Adapter 具体架构如图 3 所示，其堆叠了多个 Perceiver 块，每个 Perceiver 块都包含了交叉注意力层、潜在空间的自注意力层和前馈网络层(FFN)。具体运作流程如下，首先将 Sentence-T5 或上一个 Perceiver 块输出的嵌入和初始化的固定数量的潜在向量输入到交叉注意力层，提取关键信息，并更新潜在向量。接下来将潜在向量输入到潜在空间的自注意力层，通过自注意力机制通过计算潜在向量之间的关系，进一步提取特征并捕获上下文信息。每次自注意力操作后，潜在向量会经过一个前馈网络(FFN)。FFN 由两层全连接层和激活函数(ReLU)组成，用于进一步增强潜在向量的表达能力。前馈网络的作用是引入非线性变换，帮助模型更好地捕获复杂特征。最后经过多层 Perceiver 块处理后，潜在向量将包含输入数据的深层次特征表示。接下来，潜在向量会经过一个输出投影层，将其映射到任务相关的输出空间。

2.2.4. 模型微调

接下来，将由推理 $R=\{r_1,r_2,\cdots,r_k\}$ 经过(3)得到的软提示向量 $P=\{p_1,p_2,\cdots,p_k\}$ 与任务特定提示 Task 拼接，组成输入 x。拼接过程定义为：

$$x=[p_1;p_2;\cdots;p_k;\text{Task}] \tag{4}$$

最终的输入 X 被输入到大语言模型 M 中进行微调。微调采用低秩分解(LoRA) [16]的改进方法 LoRAMoE，仅对模型特定层进行参数高效适配。LoRAMoE 使用多个 LoRA 作为专家，并通过路由器(Router)动态控制专家的选择。此方法不仅可以在少资源的情况下达成高效训练，还能与软提示紧密结合，从而进一步提升模型性能。对于模型中的权重矩阵 $W\in\mathbb{R}^{m\times n}$，LoRAMoE 的每个专家都将其分解为两个低秩矩阵 $A\in\mathbb{R}^{m\times r}$ 和 $B\in\mathbb{R}^{r\times n}$，其中 $r<<\min(m,n)$ 是低秩维度。专家 LoRA 的权重 w_i 表示为：

$$w_i=\text{Softmax}(xw_g) \tag{5}$$

其中 w_g 为路由器(Router)的可训练权重。LoRAMoE 层取代传统 FFN 层的正向过程可以表示为：

$$W'(x)=w_0+\alpha\sum_{i=1}^{n}w_i\cdot B_iA_i \tag{6}$$

其中 w_0 是骨干模型的参数矩阵，α 是常数超参数，近似等于学习率。更新权重矩阵后的模型通过自回归生成得到最终答案，每一步的概率分布为：

$$p_\theta(a_m|a_{<m},x)=\text{Softmax}(w_{\text{head}}\cdot h+b) \tag{7}$$

其中 h 由更新后的权重 $W'(x)$ 计算得到，w_{head} 和 b 为分类头的参数，a_m 表示第 m 个词，$a_{<m}$ 表示前 $m-1$ 个词。模型最终通过最小化生成答案 a 与真实答案 a^* 之间的交叉熵损失进行优化，损失函数如下：

$$L=-E_{(q,a^*)\sim D}\left[\sum_{m=1}^{M}\log p_\theta(a_m^*|a_{<m},P(q))\right] \tag{8}$$

3. 实验设计

3.1. 数据集

为全面验证该方法的适用性和有效性，实验选用了三个知识密集型数据集，包括三个短形式生成任务和一个长形式生成任务，以评估 SOFTRAG 在多样化任务中的表现。下面是数据集的介绍。

3.1.1. 短形式生成任务

PopQA [17]：一个开放域问答数据集，旨在评估模型对流行文化知识的掌握能力。数据集涵盖电影、电视、音乐和名人等领域的问题，以问答对形式呈现，问题简短但涉及知识点广泛。答案通常为简短的实体或短语，例如人名、地名和作品名，且可能包含多个正确答案。

TriviaQA [18]：一个大规模问答数据集，包含约九万个基于事实的高难度问答对。问题来源于 trivia 风格的知识竞赛，答案可从 Web 或 Wikipedia 等资源中提取。该数据集的问题通常复杂，涉及多个知识点的组合，需要多步推理，答案为简短实体或短语。

PubHealth [19]：一个专注于生物医学领域的开放域问答数据集，旨在评估模型在生物医学知识检索和生成任务中的能力。该数据集包含了丰富的领域知识，问题来源于 PubMed 等高质量生物医学文献，涵盖药理学、疾病研究、生物技术等多个子领域，特别适合测试模型在专业领域知识上的表现。

3.1.2. 长形式生成任务

ASQA [20]：首个关注模糊事实问题的长形式问答数据集，与传统长形式答案数据集不同，每个问题

均标注了长形式答案及其对应的提取式问答对，这些问答对应能够由生成的段落回答。问题和答案均来源于 Wikipedia 文章，数据丰富且多样性强。

3.2. 检索设置

实验选用 Wikipedia 作为外部知识库，并测试多种稀疏与密集检索器，同时调整不同检索文档数量(Top-K)，以评估检索质量对模型性能的影响。测试的两种密集检索器如下：

Contriever [21]：一种基于对比学习的密集检索器，适用于高精度检索任务。对 PopQA 和 TriviaQA 这类短形式生成任务，从 Wikipedia 中检索与问题相关的文档时，Contriever 能够提供高质量的检索结果。

GTR [22]：一种基于 T5 模型的检索器，通过生成查询表示，在知识库中检索最相关的文档。对于 ASQA 的长形式生成任务，GTR 能够灵活处理多答案和模糊问题，并提供更准确的检索结果。

3.3. 评估指标

为全面衡量模型性能，实验采用多种评估指标。在 ASQA 任务中，使用官方定义的准确性(str-em)、正确性(rg)和流畅度(mau) [23]指标。对于其他任务，采用准确率评估生成答案是否包含真实答案。此外，引入了 LLM-as-a-judge [24] (GPT-4o)作为评估工具，考虑语义等价性以提供更公平的评估结果。

3.4. 基线

实验以指令调优的 Llama3-Instruct8B [25]作为骨干模型，并对比多种基线模型，具体包括以下两类：

无检索的基线模型：评估了多种强大的公开预训练大模型，包括 Llama2 7B 和 13B [26]、Llama3-Instruct8B [25]、ChatGPT [27]、SAIL-7B [28]以及指令调优模型 Alpaca-7B 和 13B [29]。对于指令调优的 LM，实验使用相应模型的官方系统提示或指令格式。这些模型仅依赖内隐的参数化知识进行生成，没有动态引入外部知识的能力。

有检索的基线模型：评估了在测试和训练阶段结合检索增强的模型，主要包括标准 RAG 基线和多种开源的大型语言模型，包括 Llama2 7B 和 13B [26]、Alpaca-7B 和 13B [29]、ChatGPT [27]。同时，还评估了一些其他的表现优异的检索增强方法，它们也根据检索到的文本进行了微调，包括 SELFRAG [8]、OPEN-RAG [17]。这些模型都使用同样的检索器根据查询以及检索到的顶级文档生成输出。

通过与这些基线模型的对比，可全面展示 SOFTRAG 在不同任务和条件下的性能优势。

4. 实验结果与分析

4.1. 与未检索的基线进行比较分析

Table 1. Overall experimental results for four tasks, categorized into retrieval and non-retrieval methods
表 1. 四项任务的总体实验结果，方法分为检索和非检索两类

	Short-form			Long-form		
	PopQA	TriviaQA	PubHealth	ASQA		
	(acc)	(acc)	acc	(em)	(rg)	(mau)
Baselines w/o Retrieval						
Llama2 7B	14.7	30.5	34.2	7.9	15.3	19.0
Alpaca 7B	23.6	54.5	49.8	18.8	29.4	61.7
Llama2 13B	14.7	38.5	29.4	7.2	12.4	16.0
Alpaca 13B	24.4	61.3	55.5	22.9	32.0	70.6

续表

ChatGPT	29.3	74.3	70.1	35.3	36.2	68.8
Llama3-Instruct8B	23.9	67.9	60.5	30.6	35.1	68.5
Baselines with retrieval						
Llama2 7B	38.2	42.5	30.0	15.2	15.2	32.0
Alpaca 7B	46.7	64.1	40.2	30.9	33.3	57.9
Llama2 13B	38.2	42.5	30.0	15.2	22.1	32.0
Alpaca 13B	46.1	66.9	51.1	34.8	36.7	56.6
Self-RAG 7B	54.9	66.1	72.0	30.2	35.7	74.9
Self-RAG 13B	56.0	67.5	76.3	31.6	35.9	69.7
RAG-ChatGPT	50.8	65.7	54.7	40.7	39.9	79.7
OPEN-RAG 7B	58.3	66.3	75.9	31.9	36.7	84.3
OPEN-RAG 13B	59.5	69.6	77.2	36.3	38.1	80.0
SOFT-RAG 8B (n = 8)	59.7	72.2	79.4	40.8	39.9	66.8

表 1 (顶部)展示了无检索基线模型的性能表现。本文提出的 SOFT-RAG 方法在所有任务中均明显优于监督微调的 LLM，甚至在部分任务和指标上超越了规模更大的模型，包括 ChatGPT。在短形式问答任务 PopQA、TriviaQA 和 PubHealth 数据集上，SOFT-RAG 8B 的准确率分别达到了 59.2%、75.2%和 79.4%，显著优于 Llama2 7B 和 Alpaca 7B 等模型。例如，在 PopQA 数据集中，SOFT-RAG 8B 的准确率几乎是 Llama2 7B (14.7%)和 Alpaca 7B (23.6%)的两倍。

在长形式问答任务 ASQA 数据集上，SOFT-RAG 8B (n = 8)在 em、rg 和 mau 指标上的得分分别为 41.8%、41.1%和 66.8%，与无检索的基线模型相比实现了大幅提升。例如，Llama2 7B 的对应得分为 7.9%、15.3%和 19.0%。这些结果表明 SOFT-RAG 在生成准确、完整且上下文相关的答案方面具有显著优势。然而，在 mau 指标上，RAG 的文本流畅度略低于原生模型，这可能是由于引入检索增强机制后对语言生成的流畅性产生了一定影响。

4.2. 与 RAG 的基线进行比较与分析

表 1 (底部)展示了与 RAG 基线的比较结果。在与 RAG 模型的对比中，SOFT-RAG 8B 继续展现出优异的性能。在短形式问答任务中，SOFT-RAG 在 PopQA、TriviaQA 和 PubHealth 数据集上的准确率分别为 59.7%、75.2%和 79.4%。相较于更大规模的 SELF-RAG 13B 和 OPEN-RAG 13B，SOFT-RAG 8B 在多数任务中表现相近或更优，同时优于其他大部分 RAG 基线模型。

在长形式问答任务 ASQA 数据集上，SOFT-RAG 8B 在 em、rg 和 mau 指标上的得分分别为 41.8%、41.1%和 66.8%。相比之下，Self-RAG 13B 的对应得分为 31.6%、35.9%和 69.7%，OPEN-RAG 13B 的得分为 36.3%、38.1%和 80.0%。虽然 SOFT-RAG 8B 在 mau 指标上略逊于 OPEN-RAG 13B，但在 em 和 rg 指标上略胜一筹，展现了在不同任务中的泛化能力与鲁棒性。这些结果进一步验证了 SOFT-RAG 在整合检索信息和生成准确答案方面的能力，同时也表明在生成文本流畅性方面仍有改进空间。

进一步分析各数据集的特性后，研究发现数据集之间在问题类型、难度以及长尾知识的覆盖上存在明显差异。以 TriviaQA 为例，其问题通常要求多步推理和跨领域知识的整合，因此 SOFTRAG 框架中检索模块与软提示适配器的协同作用显得尤为关键，能够有效弥补传统模型在复杂推理过程中的不足，进

而实现显著的性能提升。相对而言，在 PopQA 数据集中，虽然模型在准确率上也有所提升，但由于问题较为简单，长尾知识的需求较低，提升幅度相对有限

4.3. 消融实验

Table 2. Ablation experiment results ("-" indicates missing data)
表 2. 消融实验结果，"-"代表无该项数据

	PopQA (acc)	Trainable Param
Training		
Soft-RAG	59.7	0.57%
w/o soft prompt	59.4	0.57%
w/o Retrieval	28.6	0.57%
w/o LoRA	65.5	100%
Test		
w/o soft prompt	54.6	-
w/o Retrieval	22.9	-
w/o LoRA	18.6	-

为了验证 SOFTRAG 框架中各组件的作用，本文设计了消融实验，通过分别移除软提示(soft prompt)、检索模块(retrieval)和 LoRA 微调模块(LoRA)，系统分析了各模块对模型性能的贡献，如表 2 所示。实验结果表明，各模块对模型性能均具有重要作用。在完整模型配置下，SOFTRAG 在 PopQA 数据集上的准确率达到 59.7%，且仅需 0.57%的可训练参数，展现了良好的性能与效率平衡。移除软提示后，训练准确率轻微下降至 59.4%，但测试准确率显著降低至 54.6%，同时输入序列长度大幅增加，导致计算资源消耗显著上升。结果表明，软提示不仅有效减少了资源占用，还提升了模型的泛化能力。当移除检索模块时，软提示被随机初始化，训练准确率下降至 28.6%，测试准确率进一步降至 22.9%。这一结果表明，检索模块在增强模型利用长尾知识的能力方面起到了关键作用。移除 LoRAMoE 后，训练准确率虽提升至 65.5%，但此时训练参数比例增加至 100%，表明全参数微调虽然在训练阶段性能较好，但泛化能力明显不足，且资源需求显著增加。此外，由于缺乏对软提示的理解支持，测试准确率显著下降至 18.6%。综合分析表明，SOFTRAG 框架通过软提示、检索模块与 LoRAMoE 的协同作用，实现了性能与效率的最佳平衡，显著提升了模型在长尾知识任务中的表现，验证了框架设计的有效性。

4.4. 软提示数量的影响

本节的实验分析了软提示数量对于 SOFTRAG 框架的性能影响，探讨如何平衡软提示数量和模型性能。图 4 展示了软提示数量短形式问答任务的影响，随着软提示数量的增加，模型的准确率呈现出先提升后趋于平稳。当软提示数量超过某一阈值，如在 TrivialQA 中 $n = 10$ 时，准确率不再显著提升，甚至可能略有下降。这是因为过多的软提示引入了冗余信息，导致模型对关键信息的聚焦能力下降。在 PopQA 和 PubHealth 数据集中提示数量对模型性能的影响更大，在 TrivialQA 数据集性能的波动更小。但都在 $n = 8$ 左右模型的性能达到最佳，既能显著提升准确率，又不会引发显著的计算开销。图 5 展示了软提示数量对长形式任务(ASQA)的影响，软提示数量的增加对精确性(em 和 eg)指标的提升较为显著，但当 n 值较大时，模型生成的长文本可能变得冗长且不够流畅，mau 指标略有下降。实验表明，n 在 5 或 6 时是一个较为平衡的选择，能够在长形式任务中兼顾精确性和流畅度。

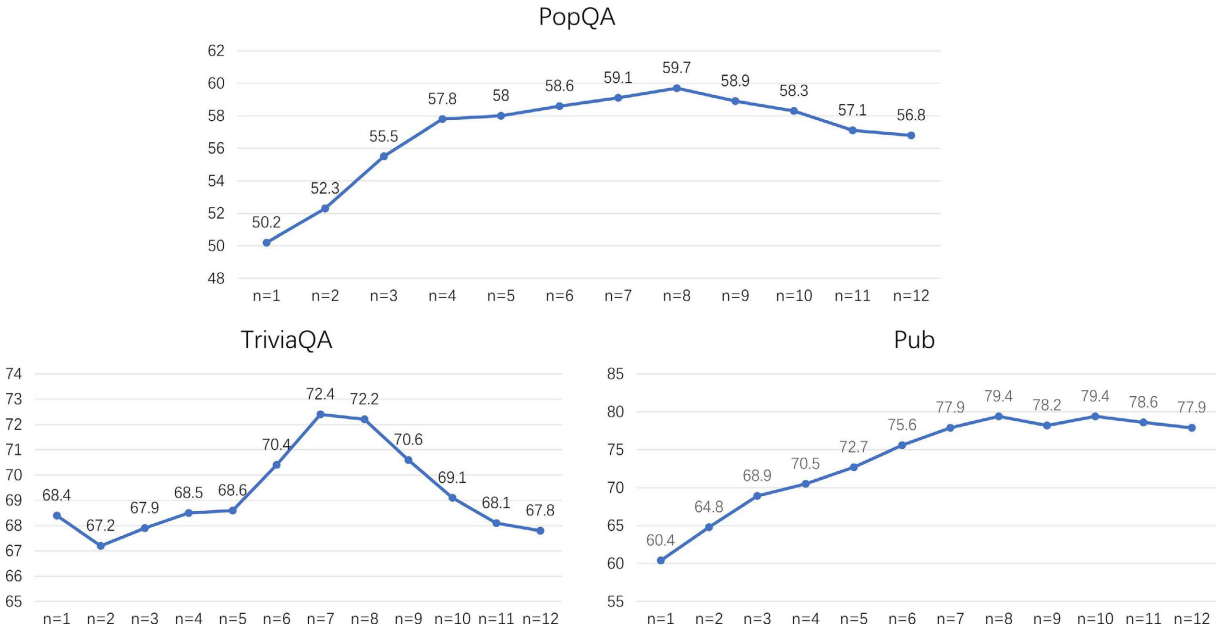

Figure 4. Line chart showing the impact of soft prompt quantity on performance in short-form question answering tasks (PopQA, TriviaQA, PubHealth)
图 4. 短形式问答任务(PopQA, TrivialQA, PubHealth)中软提示数量对性能影响折线图

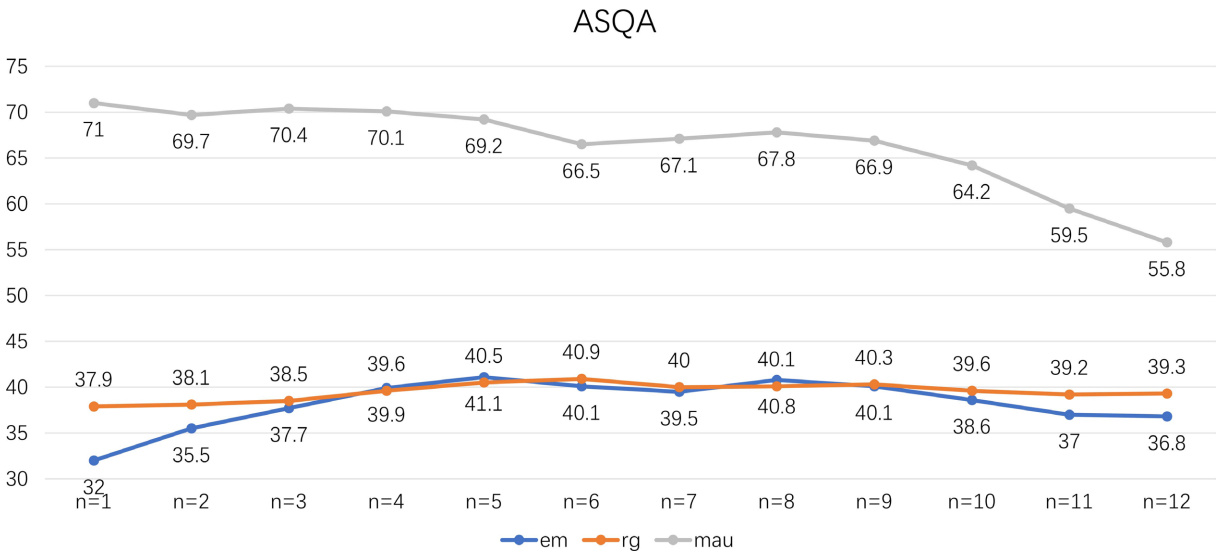

Figure 5. Line chart showing the impact of soft prompt quantity on performance in long-form question answering task (ASQA)
图 5. 长形式问答任务(ASQA)中软提示数量对性能影响折线图

4.5. 误差分析

Table 3. Error type distribution in the ASQA task
表 3. ASQA 任务误差类型分布(n = 100)

错误类型	占比	典型案例
多跳推理失败	38%	问题：“某抗生素的耐药性机制与其化学结构有何关联？” 模型未能串联“结构－靶点－耐药性”的逻辑链，仅分别描述两者。

续表

模糊问题处理不足	25%	问题：“气候变化对农业的影响有哪些？” 答案遗漏区域性差异(如热带 vs.温带)，泛泛而谈。
专业术语误解	20%	问题：“CRISPR-Cas9 的脱靶效应如何检测？” 将“脱靶”误解释为“靶向效率低”，而非非特异性编辑。
检索噪声干扰	12%	问题：“2023 年诺贝尔物理学奖得主的研究领域是什么？” 检索到过时文档(2021 年奖项)，导致答案错误。
流畅度不足	5%	答案包含重复句式或冗余连接词(如“此外，另外，同时”)。

为深入理解模型局限，本文对 ASQA 长形式任务中 SOFTRAG 的错误案例(取 100 条)进行人工标注与归类，结果如表 3 所示。结果显示，模型主要在多跳推理、模糊问题处理等方面表现不佳，说明仍需改进其动态推理能力，引导模型做出更加细化的回答。

5. 结论

本文提出了一种新型的开放域问答框架——SOFTRAG，旨在解决大语言模型在长尾知识任务中面临的挑战。通过将检索增强生成与软提示和 LoRAMoE 相结合，该框架有效平衡了性能和资源开销，显著提升了模型在短形式和长形式生成任务中的表现。SOFTRAG 框架展示了在开放域问答任务中强大的潜力，为进一步优化 LLM 在长尾知识任务中的应用提供了新的思路。然而，SOFTRAG 框架仍有改进空间。长文本生成的流畅度表现相对较弱，未来可进一步优化生成质量。此外，检索模块可能引入不相关信息，对去噪机制的优化也有待深入研究。

致　谢

感谢导师李娜在课题方向与实验设计中的专业指导与宝贵建议。同时谨向家人与朋友在项目期间的鼓励与支持致以诚挚谢意。

参考文献

[1] Floridi, L. and Chiriatti, M. (2020) GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, **30**, 681-694. https://doi.org/10.1007/s11023-020-09548-1

[2] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2019) Bert: Pretraining of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Minneapolis, 2-7 June 2019, 4171-4186.

[3] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R. and Nanayakkara, S. (2023) Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, **11**, 1-17. https://doi.org/10.1162/tacl_a_00530

[4] Kandpal, N., Deng, H., Roberts, A., Wallace, E. and Raffel, C. (2023) Large Language Models Struggle to Learn Long-Tail Knowledge. *International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 15696-15707.

[5] Zhang, T., Wang, C., Hu, N., Qiu, M., Tang, C., He, X., *et al.* (2022) DKPLM: Decomposable Knowledge-Enhanced Pre-Trained Language Model for Natural Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 11703-11711. https://doi.org/10.1609/aaai.v36i10.21425

[6] Li, D., Yan, J., Zhang, T., Wang, C., He, X., Huang, L., *et al.* (2024) On the Role of Long-Tail Knowledge in Retrieval Augmented Large Language Models. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), Bangkok, 11-16 August 2024, 120-126. https://doi.org/10.18653/v1/2024.acl-short.12

[7] Islam, S.B., Rahman, M.A., Hossain, K.S.M.T., Hoque, E., Joty, S. and Parvez, M.R. (2024) Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models. *Findings of the Association for Computational Linguistics*: *EMNLP* 2024, Miami, November 2024, 14231-14244.

https://doi.org/10.18653/v1/2024.findings-emnlp.831

[8] Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H. (2023) Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv: 2310.11511.

[9] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. and Neubig, G. (2023) Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, **55**, 1-35. https://doi.org/10.1145/3560815

[10] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z. and Zhang, Y. (2024) A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *High-Confidence Computing*, **4**, Article ID: 100211. https://doi.org/10.1016/j.hcc.2024.100211

[11] Qin, G. and Eisner, J. (2021) Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 6-11 June 2021, 5203-5212. https://doi.org/10.18653/v1/2021.naacl-main.410

[12] Dou, S., Zhou, E., Liu, Y., Gao, S., Shen, W., Xiong, L., *et al.* (2024) LoRAMoE: Alleviating World Knowledge Forgetting in Large Language Models via Moe-Style Plugin. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Bangkok, 11-16 August 2024, 1932-1945. https://doi.org/10.18653/v1/2024.acl-long.106

[13] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., *et al.* (2023) Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nature Machine Intelligence*, **5**, 220-235. https://doi.org/10.1038/s42256-023-00626-4

[14] Ni, J., Hernandez Abrego, G., Constant, N., Ma, J., Hall, K., Cer, D., *et al.* (2022) Sentence-T5: Scalable Sentence Encoders from Pre-Trained Text-To-Text Models. *Findings of the Association for Computational Linguistics*: *ACL* 2022, Dublin, 22-27 May 2022, 1864-1874. https://doi.org/10.18653/v1/2022.findings-acl.146

[15] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A. and Carreira, J. (2021) Perceiver: General Perception with Iterative Attention. *International Conference on Machine Learning*, 18-24 July 2021, 4651-4664.

[16] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W., *et al.* (2022) LoRA: Low-Rank Adaptation of Large Language Models. arXiv: 2106.09685.

[17] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. and Hajishirzi, H. (2023) When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. *Proceedings of the* 61*st Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Toronto, 9-14 July 2023, 9802-9822. https://doi.org/10.18653/v1/2023.acl-long.546

[18] Joshi, M., Choi, E., Weld, D. and Zettlemoyer, L. (2017) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, July 2017, 1601-1611. https://doi.org/10.18653/v1/p17-1147

[19] Kotonya, N. and Toni, F. (2020) Explainable Automated Fact-Checking for Public Health Claims. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 16-20 November 2020, 7740-7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[20] Stelmakh, I., Luan, Y., Dhingra, B. and Chang, M. (2022) ASQA: Factoid Questions Meet Long-Form Answers. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, 8273-8288. https://doi.org/10.18653/v1/2022.emnlp-main.566

[21] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A. and Grave, E. (2021) Unsupervised Dense in-Formation Retrieval with Contrastive Learning. arXiv: 2112.09118.

[22] Ni, J., Qu, C., Lu, J., Dai, Z., Hernandez Abrego, G., Ma, J., *et al.* (2022) Large Dual Encoders Are Generalizable Retrievers. *Proceedings of the* 2022 *Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, 7-11 December 2022, 9844-9855. https://doi.org/10.18653/v1/2022.emnlp-main.669

[23] Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y. and Harchaoui, Z. (2021) Mauve: Measuring the Gap between Neural Text and Human Text Using Divergence Frontiers. *Advances in Neural Information Processing Systems*, **34**, 4816-4828.

[24] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Stoica, I., *et al.* (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, **36**, 46595-46623.

[25] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Lample, G., *et al.* (2023) LLaMA: Open and Efficient Foundation Language Models. arXiv: 2302.13971.

[26] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Scialom, T., *et al.* (2023) LLaMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv: 2307.09288.

[27] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Lowe, R., *et al.* (2022) Training Language

Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, **35**, 27730-27744.

[28] Luo, H., Zhang, T., Chuang, Y., Gong, Y., Kim, Y., Wu, X., *et al.* (2023) Search Augmented Instruction Learning. *Findings of the Association for Computational Linguistics*: *EMNLP* 2023, Singapore, December 2023, 3717-3729. https://doi.org/10.18653/v1/2023.findings-emnlp.242

[29] Dubois, Y., Li, C. X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Hashimoto, T.B., *et al*. (2023) AlpacaFarm: A Simulation Framework for Methods That Learn from Human Feedback. *Advances in Neural Information Processing Systems*, **36**, 30039-30069.