基于强化学习的大语言模型检索增强方法

罗学炜[1] 霍雨佳[1*]
LUO Xuewei HUO Yujia

摘 要 大型语言模型（LLM）不可避免地存在幻觉现象，因其生成文本的准确性无法仅依靠模型内部封装的知识得到保证。虽然检索增强生成（RAG）为大型语言模型提供了有效的知识补充，但其效果很大程度上依赖于检索文档的相关性，当检索到无关内容时可能引发问题。为解决这一局限性，文章提出一种基于强化学习的检索增强生成框架（RL-RAG），以提升文本生成的鲁棒性。具体而言，采用交叉编码器计算的相关性分数作为奖励信号，通过建立奖励机制提升整体检索文档质量。在 3 个数据集上的实验结果表明，RL-RAG 显著提升了基于 RAG 方法的性能表现。

关键词 深度学习；自然语言处理；大语言模型；强化学习

doi：10.3969/j.issn.1672-9528.2026.02.034

0 引言

近年来，大型语言模型在各个领域引发了广泛关注，其在指令理解与生成流畅连贯文本方面展现出卓越能力[1]。然而，由于低智能度的语言识别器难以处理事实准确性[2]，且无法保证生成内容的精确性[3]，这类模型在使用过程中容易出现“幻觉”错误[4]。

先前研究已引入基于检索的技术，将相关知识整合至输入中以提升文本生成质量，例如检索增强生成。该范式通过从外部知识库检索相关文档来丰富模型输入，使 RAG 成为大型语言模型的有力补充。然而，其有效性高度依赖于检索文档的精确性与相关性。当检索过程失败或返回错误结果时，会引发对模型过度依赖检索知识的担忧。低质量检索会衍生诸多问题，包括产生大量无关信息，进而误导大型语言模型获取错误知识并生成幻觉内容。传统 RAG 的效能受限于两个关键因素：首先，其合并所有检索文档（无论相关性）会引入噪声并导致事实错误。

检索增强生成被视为解决上述问题的有效方案[5]，其通过为生成式语言模型的输入查询附加检索文档来增强性能。该方法已被证明能够从特定语料库中提供补充性知识源，从而显著提升语言模型在多项任务（尤其是知识密集型任务）中的表现。然而，这些方法往往未能解决一个关键问题：当检索失败时，后续处理机制如何应对，引文检索的根本目的是确保生成式语言模型能够获取相关且准确的知识。若检索文档被判定为不相关，检索系统反而可能加剧语言生成器产生的事实性错误。

针对前述问题，近年来，有学者在原始 RAG 框架[6]基础上发展了多种先进方法。研究认识到部分查询可能无需检索操作，反之在某些情况下，无检索的响应反而更具准确性。Asai 等[7]由此提出 Self-RAG 系统，其通过选择性检索知识并引入批判模型来决定检索必要性。Yoran 团队[8]开发了自然语言推理模型，旨在识别无关上下文并增强系统鲁棒性。SAIL 方案[9]则对指令结构进行调整，将检索文档前置嵌入指令序列。

针对上述挑战，本文重点解决了检索文本不准确的问题。提出的强化学习检索增强生成框架（RL-RAG）通过提升检索文本的相关性来改进检索效能。在该框架中，交叉编码器提供高质量的相关性评估，而强化学习则通过与环境的交互来优化检索策略。这种增强方法旨在扩展检索信息的覆盖范围。

与既往主要将检索视为增强生成模型手段的研究不同，本研究聚焦于改进检索过程中查询语句与外部知识间的语义关联度。该方法突破了传统方案依赖固定规则或单一相关性指标的局限，在提升输出可信度的同时，实现了检索质量与效率的协同优化。

1 模型描述

本文提出了一种创新性方法，通过深度融合交叉编码器与强化学习框架，引入大规模网络搜索扩展机制，以自适应方式优化文本生成的质量与可靠性。该方法在检索与生成任务中融合了两大核心技术突破：首先，采用多目标强化学习

1. 贵州民族大学数据科学与信息工程学院 贵州贵阳 550025
[基金项目] 贵州省教育厅（黔教技〔2022〕47 号）（贵州省高等学校智慧教育工程研究中心、贵州思索电子有限公司支持完成）

策略，动态调整相关性、多样性与效率的权重指标，实现检索结果的精准优化；其次，通过集成大规模网络搜索引擎，显著扩展检索范围与提高实时性，确保知识来源的广度与时效性。具体流程如图 1 所示。

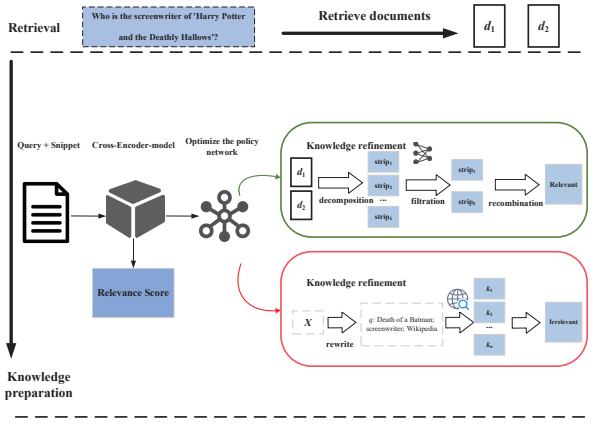

图 1 基于强化学习的检索流程图

在本节中，提出了一种优化框架，将交叉编码器（Cross-Encoder）和强化学习（Reinforcement Learning）相结合，以增强检索过程。如图 2 所示，与传统的检索增强过程不同，本研究将强化学习应用于关键检索策略，包括相关性调整、多样性控制和效率优化，从而实现了动态且智能的检索结果优化。接下来的章节将详细介绍该框架的核心组件及其背后的原理。

1.1 状态表示

状态 s 通过双编码器（Bi-Encoder）与交叉编码器（Cross-Encoder）的多层级特征融合构建。

$$s=\text{Concat}(\phi_{\text{bi}}(q),\phi_{\text{bi}}(d),\phi_{\text{cross}}(q,d)) \tag{1}$$

式中：$\phi_{\text{bi}}(q)$ 和 $\phi_{\text{bi}}(d)$ 分别表示查询 q 和文档 d 经双编码器提取的向量，其计算定义为：$\phi_{\text{bi}}(x)$ 表示输入 x 经双编码器编码并池化后的表征。

$$\phi_{\text{cross}}(q,d)=\sigma(\text{Encoder}_{\text{cross}}(q,d)) \tag{2}$$

式中：$\phi_{\text{cross}}(q, d)$ 表示查询 q 和文档 d 经交叉编码器联合编码并应用 Sigmoid 激活函数后的表征；$\text{Encoder}_{\text{cross}}(q, d)$ 表示交叉编码器对查询 q 和文档 d 的联合编码结果。

1.2 动作空间

动作 α 定义为检索结果的动态调整策略，包含以下三个组件：

（1）相关性调整：基于交叉编码器分数动态校准，计算定义为：

$$\alpha_{\text{rel}}=\frac{s_{\text{cross}}-\min(D_{\text{cross}})}{\max(D_{\text{cross}})-\min(D_{\text{cross}})+\epsilon}\times(t_{\max}-t_{\min})+t_{\min} \tag{3}$$

式中：α_{rel} 表示归一化相关性分数；s_{cross} 表示交叉编码器生成的原始分数；D_{cross} 表示交叉编码器生成的所有原始分数集合；$\max(D_{\text{cross}})$ 和 $\min(D_{\text{cross}})$ 分别表示集合中的最大值与最小值；ϵ 表示平滑因子。

冗余性控制：通过余弦相似度惩罚机制实现：

$$\alpha_{\text{div}}=\alpha_d\cdot\max(\text{sim}_{\cos}(e_d,e)),\forall e\in E_{\text{selected}} \tag{4}$$

式中：α_{div} 表示多样性分数，用于衡量当前文档与已选文档集之间的冗余度；α_d 表示控制多样性分数的强度。具体来说，计算当前文档嵌入 e_d 与已选集中各文档的最大余弦相似度，并乘以权重 α_d。α_d 越大，系统越倾向于选择与已有结果差异更大的文档。

效率优化为：

$$\alpha_{\text{eff}}=1/(\Delta t+1) \tag{5}$$

式中：Δt 表示处理时间。

动作向量 $\boldsymbol{\alpha}$ 由三个组件构成：归一化相关性分数 α_{rel}、多样性分数 α_{div} 和效率因子 α_{eff}。这些值用于动态调整检索策略，平衡相关性、多样性与计算成本。

1.3 策略网络

在策略网络中，动作向量 $\boldsymbol{\alpha}$ 的选择依赖于状态 s 的高效特征提取与动态优化。为实现该目标，设计了包含多头注意力机制（MHA）、残差连接、层归一化（LN）和多层感知机（MLP）的深度神经网络架构。这些组件的协同作用提升了模型从复杂状态空间中高效提取特征的能力，从而确保动作选择的精确性。该设计显著增强了策略网络在动态优化过程中的表达能力，使其能在复杂任务中高效决策：

$$\pi_\theta(\boldsymbol{\alpha}\mid s)=W_3\cdot\text{GeLU}(W_2\cdot\text{GeLU}(W_1\cdot\text{LN}(s+\text{MHA}(s,s,s))+b_1)+b_2)+b_3 \tag{6}$$

1.4 奖励函数

强化学习的目标是通过最大化奖励来优化策略。本任务中，奖励基于模型输出的相似度分数与实际相关性、多样性及效率的匹配程度设定：

$$R(s,\boldsymbol{a})=\alpha_r\cdot a_{\text{rel}}+\alpha_d\cdot(1-a_{\text{div}})+\alpha_e\cdot a_{\text{eff}} \tag{7}$$

式中：奖励函数 $R(s,\boldsymbol{a})$ 中的权重对指导智能体学习至关重要。

根据目标优先级初始化权重：相关性（α_r）为首要目标，其次为多样性（α_d）和效率（α_e）。最终权重通过在 PopQA 验证集上的网格搜索确定，最优组合为 α_r=0.7，α_d=0.2，α_e=0.1，该组合用于所有后续实验。

奖励函数通过综合考虑相关性、多样性和效率三个目标，为策略网络优化提供明确方向。该设计具备多目标权衡、归一化、可扩展性和任务一致性等优势，共同提升了检索与生成任务的性能。

1.5 训练过程

本研究采用近端策略优化（PPO）算法，旨在强化学习框架内实现端到端联合训练。

首先，基于双编码器和交叉编码器构建状态表示 s。双编码器提取查询和文档的独立语义嵌入，交叉编码器输出细粒度语义匹配分数。策略网络 $\pi_\theta(a/s)$ 根据当前状态生成动作向量 a，动态调整相关性权重 α_r、多样性惩罚 α_d 和效率系数 α_e。

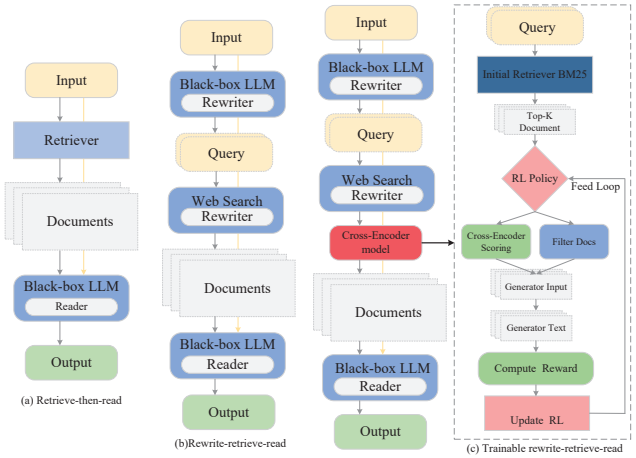

图 2 RL-RAG 算法概述

在每轮训练迭代中，系统与环境交互生成轨迹 τ，并基于多目标奖励函数计算累积奖励。随后，通过策略梯度上升更新网络参数 θ，以最大化期望回 $J_{(\theta)}$。

该框架使模型能够自适应平衡生成质量与计算效率，同时通过端到端优化避免结果冗余。详细流程见算法图。使用 trl 库实现的 PPO 算法，学习率设为 1×10^{-5}，折扣因子 γ=0.99，GAE 参数 λ=0.95，在训练集上训练 5 轮。

优势函数 A_t 通过广义优势估计（GAE）方法计算，以平衡偏差与方差：

$$A_t^{\mathrm{GAE}(\gamma,\lambda)}=\sum_{l=0}^{\infty}(\gamma\lambda)^l\delta t+l \tag{8}$$

式中：δt 为时序差分（TD）误差；γ 表示折扣因子；λ 表示控制偏差 - 方差权衡的超参数。

$$\delta t=r_t+\gamma V(s_t+1)-V(s_t) \tag{9}$$

式中：$V(s_t)$ 表示状态价值函数。

2 实验与结果分析

2.1 数据集

RL-RAG 在三个数据集上进行了评估：PopQA [10]、PubHealth [11] 和 Arc-Challenge[12]。

在实验中使用了 NVIDIA A800 80 GB GPU。对于 LLaMA-2（7B）的生成，在推理过程中会占用超过 40 GB 的内存。

2.2 基线模型

主要比较了两种 RL-RAG 的方法，一种是带有检索的，另一种是不带检索的。其中，不带检索的方法可以进一步细分为标准 RAG 和最新的高级 RAG。

不带检索的基线模型。评估了几种公开的大语言模型，包括 LLaMA2-7B、指令微调模型和 CoVE65B。此外，还包含了专有的大语言模型，如 LLaMA2-chat13B 和 ChatGPT。

标准 RAG。评估了标准的 RAG 模型，其中使用了几种公开的指令微调大语言模型，包括 LLaMA2-7B、LLaMA2-13B、Alpaca-7B、Alpaca-13B，以及 Self-RAG 中指令微调的 LLaMA2-7B。

高级 RAG。Self-RAG 对 LLaMA2 进行了指令微调，包括几组由 GPT-4 标记的反思标记。此外，还参考了使用专有数据训练的检索增强基线的结果。并引入了 CRAG 生成作为强大的基线进行比较。CRAG 的核心思想是对检索到的文档进行质量评估和自我修正。CRAG 是一个轻量级的、即插即用的修正框架，无需额外的指令微调即可运行。

2.3 评价指标

按照以往的研究，使用准确率作为 PopQA、PubHealth 和 Arc-Challenge 的评估指标。

2.4 实验细节

为确保公平比较，所有基线方法和 RL-RAG 均使用相同的检索器（DPR）和相同的交叉编码器（roberta-large-mnli）进行相关性评分。还使用了相同的提示模板，并将所有模型的上下文窗口限制为 2 048 个标记，以保持输入长度的一致性。

进行实验，以证明 RL-RAG 对基于 RAG 的方法的适应性及其在生成任务中的通用性。所有超参数，包括奖励权重（α_r=0.7，α_d=0.2，α_e=0.1），通过在 PopQA 的验证集上进行网格搜索单独选择的。这些值在随后对 PubHealth 和 Arc-Challenge 的评估中保持固定，没有进一步调整。

2.5 整体性能

在对 PopQA、PubHealth 和 Arc-Challenge 数据集的评估中，与传统的 RAG 相比，RL-RAG 展现出显著的性能提升。结果总结在表 1 中。例如，与 SelfRAG-LLaMA2-7B 相比，RL-RAG 在 PopQA 上的准确率高出 6.1%，在 PubHealth 上的准确率高出 35.1%。系统的效果进一步通过在 Arc-Challenge 上 16.7% 的准确率提升得到证实。当底层模型转变为 LLaMA2-7B 时，RL-RAG 仍然优于 RAG，在 PopQA 上领先 4.9%，在 PubHealth 上领先 16.5%。这一结果证明了 RL-RAG 的有效性和稳健性。与最新的 CRAG 相比，RL-RAG 在 PopQA、PubHealth 和 Arc-Challenge 上分别提高了 0.5%、1.9% 和 2.8%。这些结果突显了 RL-RAG 在不同情境下的有效性及其在实际应用中的潜力。

表 1 PopQA、PubHealth 和 ARC 数据集测试集的整体评估结果

Method	PopQA/%	PubHealth/%	ARC/%
LMs trained with propriety data			
LLaMA2-13B	20.0	49.4	38.4
Ret-LLaMA2-13B	51.8	52.1	37.9
ChatGPT	29.3	70.1	75.3
Ret-ChatGPT	50.8	54.7	75.3

表 1（续）

Method	PopQA/%	PubHealth/%	ARC/%
Baselines without retrieval			
LLaMA2-7B	14.7	34.2	21.8
Alpaca-7B	23.6	49.8	45.0
LLaMA2-13B	14.7	29.4	29.4
Alpaca-13B	24.4	55.5	54.9
Baselines with retrieval			
LLaMA2-7B	38.2	30.0	48.0
Alpaca-7B	46.7	40.2	48.0
SAIL	—	69.2	48.4
LLaMA2-13B	45.7	30.2	26.0
Alpaca-13B	46.1	51.1	57.6
LLaMA2-hf-7b			
RAG	50.5	48.9	43.4
Self-RAG	29.0	0.7	23.9
CRAG	54.9	59.5	53.7
RL-RAG (Ours)	55.4	61.4	56.5
SelfRAG-LLaMA2-7b			
RAG	52.8	39.0	53.2
Self-RAG	54.9	72.4	67.3
CRAG	59.8	74.1	68.6
RL-RAG (Ours)	59.9	75.8	68.9

3 结语

本文针对基于 RAG 方法在检索相关性不足时导致生成知识不准确的问题，提出了一种基于强化学习的检索方法（RL-RAG）。该方法通过优化知识利用并结合网络搜索，显著提升了自动自我修正能力和检索文档的高效利用。实验结果表明，RL-RAG 在多种生成任务中表现出良好的适应性和效率，增强了生成模型的自我修正能力并优化了知识利用效率，从而在动态环境中实现更准确的生成决策。尽管如此，RL-RAG 仍有改进空间，未来需进一步探索其在不同领域和任务中的泛化能力，以及开发更强大的内置评估能力以实现完全自适应的生成和知识检索过程。

参考文献：

[1]BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL].(2020-07-22)[2025-12-30]. https://doi.org/10.48550/arXiv.2005.14165.

[2]MIN S , LYU X X , YIH W T ,et al. FActScore: fine-grained atomic evaluation of factual precision in long form text generation[EB/OL].(2023-11-22)[2025-11-05].https://arxiv.org/abs/2305.14251.

[3]MUHLGAY D, RAM O, MAGAR I, et al. Generating benchmarks for factuality evaluation of language models[C]// Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics.Brussels: ACL, 2024:49-66.

[4]JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J].ACM computing surveys,2023, 55(12): 1-38 .

[5]GUU K, LEE K, TUNG Z, et al. REALM: retrieval augmented language model pre-training[EB/OL].(2020-02-10)[2025-12-25].https://doi.org/10.48550/arXiv.2002.08909.

[6]KIM J, NAM J, MO S, et al. SuRe: summarizing retrievals using answer candidates for open-domain QA of LLMs[EB/OL].(2024-04-17)[2025-12-25].https://doi.org/10.48550/arXiv.2404.13081.

[7]ASAI A, WU Z Q, WANG Y Z, et al.Self-RAG: learning to retrieve, generate, and critique through self-reflection[EB/OL].(2023-10-17)[2025-12-25].https://doi.org/10.48550/arXiv.2310.11511.

[8]YORAN O, WOLFSON T, RAM O, et al. Making retrieval-augmented language models robust to irrelevant context[EB/OL].(2024-05-05)[2025-12-25].https://doi.org/10.48550/arXiv.2310.01558.

[9]LUO H Y, ZHANG T H,CHUANG Y S, et al. SAIL: search-augmented instruction learning[C]//Findings of the Association for Computational Linguistics: EMNLP 2023. Brussels:ACL,2023:3717–3729.

[10]MALLEN A, ASAI A, ZHONG V, et al.When not to trust language models: Investigating effectiveness of parametric and non-parametric memories[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics.Brussels:ACL,2023:9802-9822.

[11]ZHANG T H, LUO H Y, CHUANG Y S, et al.Interpretable unified language checking[EB/OL].(2023-04-07)[2025-12-02].https://doi.org/10.48550/arXiv.2304.03728.

[12]BHAKTHAVATSALAM S, KHASHABI D, KHOT T, et al.Think you have solved direct-answer question answering? try ARC-DA, the direct-answer AI2 reasoning challenge[EB/OL].(2021-02-05)[2025-12-29].https://doi.org/10.48550/arXiv.2102.03315.

【作者简介】

罗学炜（2001—），男，福建三明人，硕士研究生，研究方向：大语言模型。

霍雨佳（1982—），通信作者（email: huo.yujia@gzmu.edu.cn），男，贵州遵义人，博士，副教授、硕士生导师，研究方向：深度学习、智慧教育、自然语言处理等。

（收稿日期：2025-10-15 修回日期：2026-02-04）