SRAS：面向边缘原生RAG流水线的轻量级强化学习文档选择器

Rajiv Chaitanya Muttur 本科生，计算机科学与工程系，
Dayananda Sagar工程学院，班加罗尔，印度 电子邮箱：
rajiv.muttur@gmail.com ORCID：0009-0007-7159-5610

arXiv:2601.01785v1 [cs.IR] 5 Jan 2026

摘要—**检索增强生成（RAG）系统通常依赖固定的 top-k 文档选择机制，忽视下游生成质量，并带来计算开销。我们提出 SRAS（稀疏奖励感知选择器），这是一种通过强化学习（RL）训练的轻量级文档选择器，适用于边缘原生RAG部署。不同于假定拥有充足内存和延迟预算的既有基于RL的检索器，SRAS 使用近端策略优化（PPO）学习紧凑的（∼0.76MB）策略，并由结合宽松 F1 与 BERTScore 的混合奖励信号提供指导。我们的方法在严格的 token 数量和计算资源约束下运行，同时在 CPU 上保持 <1s 的延迟。SRAS 在一个合成问答基准上优于监督式和随机选择器，并能泛化至真实数据；无需针对特定领域进行调优，即可在 SQuAD v2 上取得 0.8546 的 BERTScore F1。本文首次证明，基于RL的文档选择可以做到超轻量、具备延迟感知能力，并适用于端侧 RAG流水线。**

索引词——**检索增强生成、强化学习、文档选择、边缘计算、近端策略优化、BERTScore、低延迟自然语言处理。**

一、引言

检索增强生成（RAG）已成为一种强大的范式，可通过外部知识源 [1], [2] 增强大语言模型（LLM）。在典型的 RAG流水线中，给定一个查询，检索器会从大规模语料库中选出最相关的前 k 个文档，再将其传递给生成器以生成最终输出。这种两阶段架构可减少幻觉，并促进生成有事实依据的响应。然而，检索步骤至关重要：如果所选文档不相关或存在冗余，即使生成器能力很强，也无法弥补。

大多数现有的 RAG系统依赖静态启发式检索，包括稀疏检索（例如 BM25）或稠密检索（例如基于句子嵌入的余弦相似度），但这些方法并未针对下游任务性能进行优化。它们不考虑生成质量，也不利用生成结果的反馈，导致检索与生成之间出现错位，在低监督场景中尤为明显。

这一挑战在边缘设备部署中尤为突出，因为计算资源、内存和功耗方面的限制要求使用紧凑且延迟低的模型 [12]。DPR [4] 或 ColBERT [5] 等大型学习型检索器因资源需求过高，不适用于这类环境。此外，传统检索方法并未利用弱监督信号（例如问题和答案的正确性）来改进文档选择。

为解决这些局限，我们提出*SRAS* （稀疏奖励感知选择器），这是一种紧凑型神经文档选择器，利用稀疏的基于问答的奖励信号进行强化学习（RL）训练。SRAS 通过学习到的非线性评分策略取代固定的 top-k 检索，该策略使用近端策略优化（PPO） [13]进行训练。其混合奖励函数结合了宽松 F1 和 BERTScore，能够捕捉生成答案与参考答案之间的词汇和语义对齐。

我们在一个采用合成监督的问答基准测试上评估 SRAS，并将其与多种强基线方法进行比较，包括基于 top-k 余弦相似度的方法、随机选择以及基于监督的注意力模型。SRAS 在保持答案准确率具有竞争力的同时，将模型大小控制在 1MB 以下，并在 CPU 上实现低于 0.6s 的推理延迟，因此非常适合在真实场景中进行边缘部署。

我们的消融研究凸显了混合奖励塑形、监督预热和课程学习在实现稳定且样本高效的强化学习训练方面的作用。通过将强化学习与检索相结合，SRAS 为构建自适应、任务感知且轻量级的检索增强生成系统开辟了新方向。

我们的主要贡献如下：

- 我们提出了 SRAS，这是一种轻量级、基于强化学习的文档选择器，能够对查询-文档交互进行建模，并从稀疏问答监督中学习。
- 我们引入了一种结合宽松 F1 和 BERTScore 的混合奖励信号，以指导学习，而无需密集标注。
- 我们证明，SRAS 在边缘设备约束下仍能实现出色的问答性能，同时模型大小低于 1 MB，推理延迟低于 1 秒。
- 我们提供了详细的消融实验，突显了奖励塑形、监督预热和课程学习对于有效策略训练的重要性。
- 我们在来自 SQuAD v2 的 250 个真实问答示例上对 SRAS 进行零样本评估，结果表明其具有出色的语义泛化能力。

II. 相关工作

A. RAG流水线中的静态检索与学习型检索

检索增强生成（RAG）架构通过从语料库中检索外部文档来增强语言模型 [1], [2]。传统方法依赖基于静态相似度的检索，采用 BM25 等稀疏方法或 Sentence-BERT 等稠密方法 [3]。这些方法效率很高，但缺乏适应性，因为它们既不利用下游任务反馈，也不针对答案质量优化检索。

为解决这一问题，研究人员提出了密集段落检索（DPR） [4] 和 ColBERT [5]等学习型检索器。这些模型通过对比学习目标进行训练，将查询和文档嵌入到共享向量空间中。尽管它们提高了检索准确率，但其较大的模型大小和计算资源需求使其不适用于低延迟或边缘场景。

B. 用于检索与重排序的强化学习

强化学习（RL）已被应用于信息检索，以根据反馈信号优化排序策略 [6], [7]。在问答系统中，强化学习被用于利用答案级反馈训练检索器，而不是使用显式的文档标签 [8], [9]。然而，大多数现有工作要么训练完整的检索器编码器，要么将检索与生成结合为单个强化学习智能体，从而导致较高的模型复杂度。

近期研究探索了利用生成输出衍生的奖励信号进行模块化训练，以改进检索[10]。然而，这些方法仍依赖大型预训练编码器，因此不适用于受限环境。

C. 资源约束下的文档选择

在边缘设备上部署RAG系统会带来独特挑战：必须在不牺牲准确性的前提下，尽可能降低内存占用、推理延迟和功耗[12]。大多数学习型检索器并不适合这类场景，因为它们需要GPU加速和大容量内存缓冲区。

针对模型压缩和神经排序，有研究提出了采用浅层架构或基于投影的评分函数的轻量级选择器 [11]。然而，将其应用于监督信号稀疏的QA驱动型RAG流水线，仍未得到充分探索。此外，直接针对边缘约束下基于强化学习的文档选择开展的研究也寥寥无几。

相比之下，我们提出了一种紧凑的前馈选择器，通过稀疏的QA衍生奖励学习选择相关文档。SRAS将检索与大型编码器解耦，并着重利用弱监督进行学习，为低资源RAG提供了一种新范式。

III. 方法论

我们提出了一种模块化、基于强化学习的检索框架，专为适用于边缘环境的检索增强生成部署而设计。该框架的核心是*SRAS*（稀疏奖励感知选择器），这是一种紧凑型神经选择器，通过下游问答性能得出的弱监督信号进行训练，以优化文档选择。本节将介绍流水线架构、奖励设计，以及使用近端策略优化（PPO）[13]的训练流程。

A. 流水线概述

我们的 SRAS增强型RAG流程由四个模块化阶段组成：

1)语料预处理和嵌入： 所有原始文档均经过预处理，转换为扁平化 JSON 结构，并使用轻量级 Sentence-BERT 编码器（MiniLM-L6-v2） [3], 进行嵌入，输出结果缓存为 .npy 和 .pt 格式，以便复用。

2) 合成问答对生成： 生成式模型（valhalla/t5-base-qg-hl）在语料库上生成合成问答对。这些问答对将作为训练样本，用于评估文档选择的有效性。

3) 通过 SRAS 进行文档选择： 给定一个问题嵌入和一组候选文档嵌入（通常为 $n = 8$），SRAS 通过轻量级前馈评分函数计算分数，该函数对查询与文档之间的交互进行建模，并选择排名前-k的文档供后续使用。

4) 答案生成和奖励计算： 将选定的文档输入冻结的问答模型（例如 T5 或 FiD）以生成答案。将生成的输出与标准答案进行比较，以计算混合奖励信号。随后使用该奖励更新 SRAS 策略。

图 1 展示了 SRAS增强型RAG流程的完整架构。

B. 混合奖励信号：基于 *QA* 的反馈

由于无法获得相关文档的真实标签，我们将 QA 答案质量视为检索效用的代理指标。因此，我们定义了一种混合奖励信号R，结合了生成答案与参考答案之间的词汇相似度和语义相似度：

$$R = \alpha \cdot \text{Relaxed-F1} + (1 - \alpha) \cdot \text{BERTScore} \tag{1}$$

我们基于使用训练问答对的留出子集对 $\alpha \in \{0.3, 0.5, 0.6, 0.7\}$ 进行的小规模网格搜索，设置了 $\alpha = 0.6$ 。从经验上看，该值在策略学习中较好地权衡了早期阶段的词汇对齐（通过宽松 F1）与稳定的梯度信号（通过 BERTScore ）。较低的 α 值会导致奖励收敛较慢，而较高的值则会造成对词元级模式的过拟合。

- **宽松 F1：** 一种基于规范化答案计算的宽松词元级 F1 分数（移除标点符号、转换为小写并去除停用词），用于衡量部分词汇重叠，并容忍轻微的措辞差异。

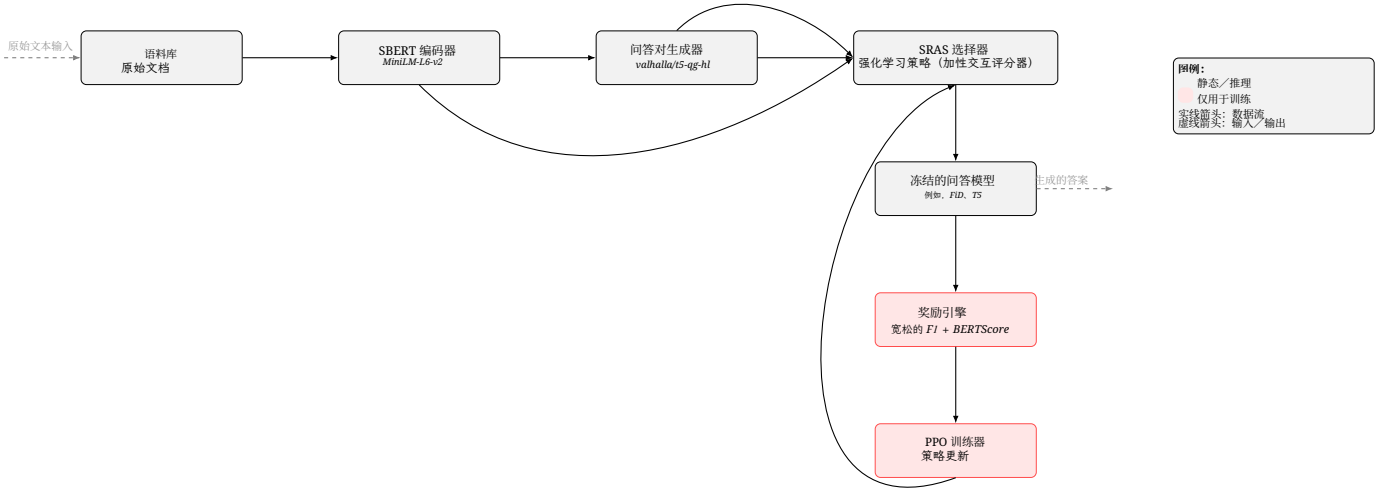


图 1. 端到端 SRAS增强型RAG流程。选择器使用来自 SBERT 和问答对生成器的嵌入。奖励引擎和近端策略优化仅用于训练。

•**BERTScore：** 一种语义相似度指标，通过冻结的 RoBERTa-large 模型生成的上下文化词元嵌入之间的余弦相似度进行计算， [14],能够捕捉更深层的含义并具备释义鲁棒性。

这种混合奖励鼓励 SRAS 选择能够生成词汇准确且语义忠实的答案的文档，同时增强对合成问答对中弱监督和冗余的鲁棒性。这种混合方式还能通过平滑早期探索期间稀疏或嘈杂的奖励信号，稳定近端策略优化训练。

*C. SRAS*架构

SRAS模型设计精简且硬件高效。给定一个问题嵌入 $q \in \mathbb{R}^d$ 和一组文档嵌入 $D = \{d_1, \ldots, d_n\} \subset \mathbb{R}^d$，我们使用以下方式为每个文档 d_i 计算相关性分数 s_i ：

$$h_q = W_q q \in \mathbb{R}^h \tag{2}$$

$$h_{d_i} = W_d d_i \in \mathbb{R}^h \tag{3}$$

$$s_i = w^\top \tanh(h_q + h_{d_i}) \in \mathbb{R} \tag{4}$$

其中， $W_q, W_d \in \mathbb{R}^{h \times d}$ 为可学习的投影矩阵， $w \in \mathbb{R}^h$ 为学习得到的注意力向量。该模型约有 197K 个参数，总大小为~0.76 MB。

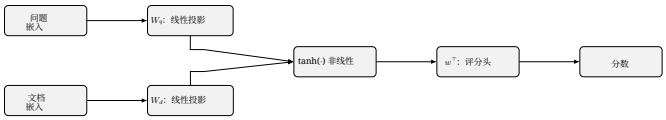


图 2。SRAS 评分架构。问题和文档嵌入被投影到共享空间，通过 tanh 非线性进行组合，并进行线性评分。

图 2 详细展示了 SRAS 评分机制。

D. 稀疏奖励下的近端策略优化训练

我们采用 PPO 算法 [13] ，根据下游 QA 衍生奖励训练 SRAS。文档选择任务被建模为在n 个候选项中进行离散动作选择问题，智能体为每个 QA 对选择 k 篇文档。

1）PPO Configuration:

- **训练轮次：** 25
- **批次大小：** 8
- **Top-k 选择：** 3
- **学习率：** 1×10^{-5}
- **折扣因子** γ**：** 0.99
- **裁剪** ϵ**：** 0.2
- **优化器：** AdamW

2) 稳定化技术： 为提升样本效率并增强稀疏反馈下的鲁棒性，我们采用：

- **监督预热：** 使用黄金标准QA文档标签进行交叉熵预训练，以初始化选择器。
- **奖励归一化：** 将每个批次归一化为零均值和单位方差。
- **优势估计：** 使用类似GAE的估计器来平滑优势值。
- **课程学习：** 我们从较容易的QA对（Top-1重叠率较高）开始训练，并逐步提高难度。

这些组件共同支持在弱监督下进行高效的近端策略优化训练，同时保持较低的推理延迟和内存使用量。

IV. 实验设置

本节概述用于在适合边缘部署的条件下评估 SRAS 选择器的数据集构建、候选采样、模型基线、评估协议和部署约束。

A. 合成问答数据集

为了在没有金标准标签的情况下实现由奖励驱动的文档选择，我们使用精心整理的、包含 905 篇不同文档的语料库构建合成问答数据集。每篇文档都转换为扁平化的 JSONL 格式，以便高效地进行索引、嵌入和问答生成。

我们采用 valhalla/t5-base-qg-hl（一种预训练模型），通过基于高亮的提示策略 [15]生成问答对。最终共生成 750 个高质量问答对，每个问答对都关联一篇金标准上下文文档。

B. 候选文档池化

对于每个问答对，我们构建一个包含 $n = 8$篇文档的候选池：1 篇金标准上下文文档，以及从语料库中随机抽取的 7 篇干扰文档。此设置模拟了噪声检索环境，同时使选择复杂度保持在资源受限推理可处理的范围内。

所有文档和查询均使用冻结的 MiniLM-L6-v2 Sentence-BERT 编码器 [3]进行编码。为提高效率，嵌入会缓存并在训练和评估期间重复使用。

C. 评估指标

我们使用面向问答和面向部署的指标评估选择器性能：

- **宽松 F1：** 经规范化（转换为小写、标点符号处理和停用词移除）后计算的词元级 F1 分数。该指标衡量与真实答案之间的部分词汇重叠。
- **BERTScore F1：** 基于 RoBERTa-large [14],的上下文嵌入计算的语义相似度分数，用于捕捉生成答案与参考答案之间更深层次的对应关系。
- **延迟：** 在 Intel i5 CPU 上以单线程模式、批次大小为 1 时测得的每次查询平均 CPU 推理时间（毫秒）。
- **模型大小：** 序列化选择器模型的大小（MB），包括权重和配置。

D. 部署约束

所有选择器均在仅 CPU 环境中进行评估，以模拟真实场景下的边缘设备推理。下游问答模型固定为 T5-small，采用贪心解码（束宽为 1），最大输出长度为 32 个词元。

推理时间以批次大小为 1 进行测量，以反映真实的单次查询延迟。为保证公平性，所有文档嵌入和问答模型在各次评估中均保持不变。

E. 基线方法

我们将 SRAS 与以下标准选择器和学习型选择器进行比较，结果汇总于表 I:

- **Top-k 余弦：** 一种静态稠密检索器，使用 MiniLM SBERT 根据文档与查询嵌入之间的余弦相似度对文档进行排序。
- **随机：** 从候选集中均匀随机采样 3 份文档，提供性能下界。
- **监督式（FF）：** 使用金标准文档标签上的交叉熵损失进行训练的前馈神经选择器。与 SRAS架构相匹配。
- **SRAS（PPO）：** 我们提出的选择器，使用混合式问答奖励通过 PPO 进行端到端训练，且无法访问文档标签。

所有模型都从包含 8 个文档的候选池中选择 $k = 3$ 个文档，并将输出传递给同一个冻结的问答模型。

表 I 选择器基线对比

选择器	评分函数	训练方法	是否经过学习?
Top-k 余弦	余弦相似度（SBERT）	None	No
随机	均匀采样	None	No
监督式（FF）	$\tanh(W_q q + W_d d_i)$	交叉熵	Yes
SRAS（PPO）	$\tanh(W_q q + W_d d_i)$	PPO	Yes

F. 原型设计选择：

为确保强化学习（RL）训练过程易于处理，并在各消融变体之间进行受控评估，我们有意将语料库限制为 100 篇文档，并采用冻结的问答生成流水线。此限制有助于实现可复现的奖励计算和公平的基准测试，尤其是在监督信号稀疏且延迟的情况下。尽管此类简化可能会限制实际应用规模，但它们使我们能够以最少的混杂因素隔离并评估策略改进（例如奖励塑形、课程学习）的影响。轻量级设置也进一步契合了我们对高效设备端检索增强生成选择器进行原型开发的目标。

V. 结果与讨论

本节在边缘设备约束下，对 SRAS 与非学习型和学习型文档选择器进行比较评估。我们使用多种定量指标和可视化结果，分析答案质量、延迟以及训练增强的作用。

A. 文档选择性能

表II报告了所有文档选择器在包含300个问答对的留出测试集上的性能。SRAS（近端策略优化基线）的宽松 F1 为 0.1473，BERTScore F1 为 0.8463；其宽松 F1 优于监督式选择器，语义相似度则与之相近。

有趣的是，Top-k 余弦基线的整体表现最佳（宽松 F1：0.1604，BERTScore F1：0.8549），但它是一种静态、不可训练的方法，完全依赖冻结的 MiniLM 嵌入。尽管它的延迟和模型占用均为零，却缺乏适应能力，也无法利用任务特定反馈。

为更清晰地展示这些权衡，图 3 给出了延迟与宽松 F1 的气泡图，其中气泡大小表示 BERTScore F1。SRAS 在答案质量与推理效率之间取得了良好平衡，准确率接近监督式方法，同时延迟低于一秒且模型大小紧凑。

表II 边缘约束下文档选择器的性能（在 300个问答对上评估）

选择器	宽松 F1	BERTScore	延迟（秒）	大小（MB）
Top-k 余弦相似度	0.1604	0.8549	0.07	0.00
随机	0.1182	0.8344	0.10	0.00
监督式（FF）	0.1323	0.8511	0.46	0.76
SRAS（PPO）	0.1473	0.8463	0.38	0.76

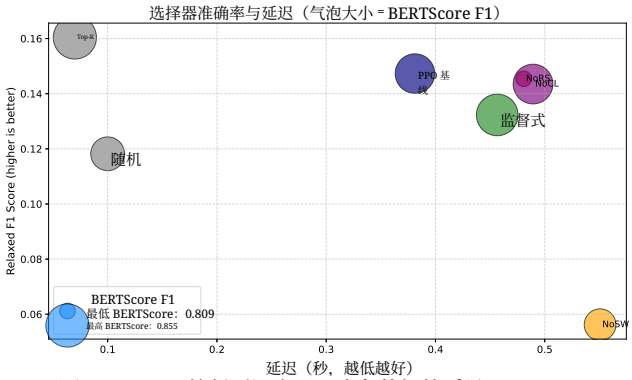


图 3. SRAS 以较低延迟实现了出色的问答质量。

B. 强化学习增强措施的影响

为单独评估各项近端策略优化训练策略的贡献，我们进行了消融研究，分别禁用了：(i) 监督预热（NoSW）、(ii) 奖励塑形（无 RS）和 (iii) 课程学习（NoCL）。图 4 展示了各变体的宽松 F1 和 BERTScore F1，并在柱形图上标注了延迟。

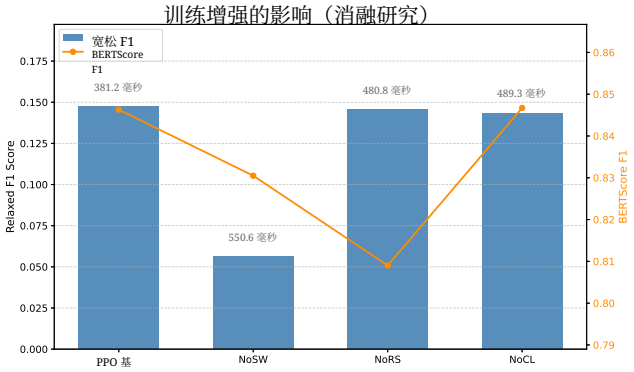


图 4. 消融研究展示了移除监督预热（SW）、奖励塑形（RS）和课程学习（CL）的影响。

主要发现：

- **奖励塑形（RS）**：移除该措施会导致宽松 F1 出现最显著的下降（0.1473 → 0.0562），这表明仅靠稀疏问答奖励不足以实现有效学习。
- **监督预热（SW）**：NoSW 会导致训练不足（BERTScore F1 降至 0.8305）和更高的延迟，凸显了从标注监督中进行自举的价值。
- **课程学习（CL）**：NoCL 会使宽松 F1 略微下降（降至 0.1435），并增加奖励方差，这表明它对提升训练稳定性的作用大于对最终性能的提升。

C. PPO 奖励收敛

图 5 展示了近端策略优化基线和消融变体在各训练轮次中的奖励变化。完整的近端策略优化智能体（包含 SW、RS 和 CL）稳定收敛在约 0.42 的奖励值。

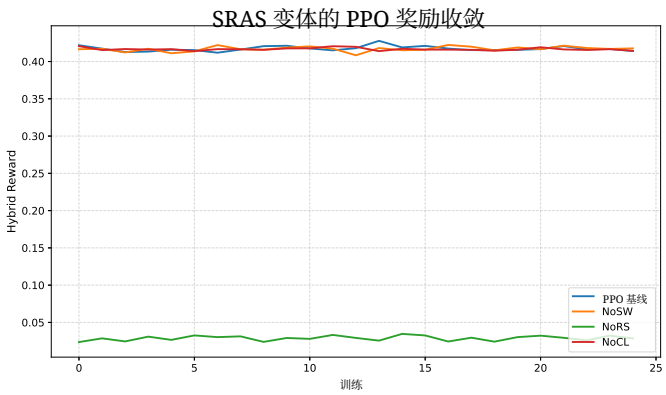


图 5. 每轮的平均近端策略优化奖励。移除 RS 会导致学习停滞。SW 和 CL 可提升训练初期的稳定性。

无 RS 无法学习，奖励停滞在约 0.02–0.03，证实了在稀疏反馈场景中塑形奖励的重要性。

NoSW 最终奖励相近，但在训练初期波动更大，凸显了监督初始化对提升稳定性的作用。

NoCL收敛表现相当，但进展不够稳定，说明课程学习提升的是训练过程的平稳性，而非最终结果。

D. 向真实世界问答的泛化：*SQuAD v2* 评估

为评估 SRAS 在真实场景下的表现，我们在来自 SQuAD v2 的 250 个问答对上进行了测试。SQuAD v2 是一个包含可回答和不可回答问题的基准测试。所有选择器（SRAS、Top-k、随机）均为每个问题选择一个文档，随后使用 T5-base 生成答案。我们使用宽松 F1 和 BERTScore F1，根据标准答案片段对答案进行评估。

表 III SQuAD V2 中 250 个问答对上的性能。

选择器	宽松 F1	BERTScore F1
SRAS（PPO 基线）	**0.0454**	**0.8546**
随机	0.0313	0.0956
Top-k 余弦	0.0256	0.1282

尽管 SRAS 是在合成数据上训练的，但它仍能有效泛化，达到最高的语义对齐度（BERTScore F1：0.8546），并优于两种基线方法。

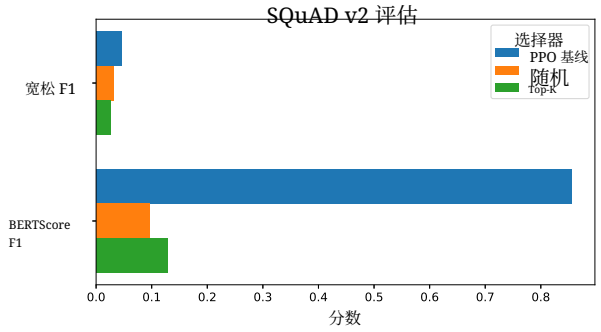


图 6. SRAS 与基线方法在 SQuAD v2 上的对比：宽松 F1 和 BERTScore F1。尽管存在领域偏移，SRAS 仍具有良好的泛化能力。

VI. 结论

我们提出了 SRAS（稀疏奖励感知选择器），这是一种轻量级的强化学习文档选择框架，适用于边缘约束下的检索增强生成（RAG）流水线。SRAS 以一个通过近端策略优化（PPO）训练、并由基于问答的任务奖励引导的紧凑策略，取代了传统的 top-k检索。与基于相似度的启发式方法或静态分类器不同，SRAS 能自适应地选择可最大限度提升下游问答性能的文档，并能很好地泛化到未见过的查询。

为应对现实问答场景中稀疏且延迟的奖励所带来的挑战，我们引入了三项关键训练增强：监督预热、奖励塑形和课程学习。消融实验表明，每一项都不可或缺：奖励塑形可加快早期学习，监督预热可稳定策略初始化，而课程学习则使扩展至更大的候选池成为可能。

在一个包含 300 个样例的问答基准测试中（使用现有语料库数据），SRAS 的宽松 F1 达到 0.1473，BERTScore F1 达到 0.8463，表现可与监督式基线方法相媲美，并接近 oracle Top-k 检索，同时保持亚秒级延迟，且模型大小不足 1MB。在 250 个未见过的 SQuAD v2 问题上进行的评估证实，SRAS 能够泛化至其训练分布之外。

这些特性使 SRAS 非常适合计算资源受限的部署。

局限性与未来工作

SRAS 基于静态抽取式问答语料库进行训练，其领域多样性有限。未来工作包括扩展到多领域和多语言语料库，拓展至生成式或抽象式问答任务，并与可微检索器集成以进行端到端训练。我们计划在嵌入式硬件上评估 SRAS，并探索通过量化和剪枝进一步压缩模型。目前，检索器与其他组件相互独立且固定；将 SRAS 集成到端到端检索器-生成器框架中仍有待后续研究。

最终，SRAS朝着在边缘原生检索增强生成系统中实现实时、高质量的文档选择迈出了一步：弥合基于学习的检索与实用端侧智能之间的差距。

致谢

作者承认，在准备稿件期间使用了 OpenAI 的 ChatGPT 进行语言润色和少量格式调整。所有研究构想、方法论、实验和结果均为作者独立完成。

参考文献

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kandpal, B. Stoyanov, and S. Riedel, “面向知识密集型自然语言处理任务的检索增强生成”，载于 神经信息处理系统进展（*NeurIPS*），第 33 卷，第 9459–9474 页，2020 年。[2] G. Izacard and E. Grave, “利用段落检索与生成模型进行开放域问答”，载于 第 *16* 届欧洲计算语言学协会分会会议论文集（*EACL*），第 874–880 页，2021 年。[3] N. Reimers and I. Gurevych, “Sentence-BERT：使用孪生 BERT 网络的句子嵌入”，载于 自然语言处理实证方法会议论文集（*EMNLP*），第 3982–3992 页，2019 年。[4] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, “面向开放域问答的密集段落检索”，载于 自然语言处理实证方法会议论文集（*EMNLP*），第 6769–6781 页，2020 年。[5] O. Khattab and M. Zaharia, “ColBERT：通过 BERT 上的上下文化后期交互实现高效且有效的段落搜索”，载于 第 *43* 届 *ACM* 国际信息检索研究与发展会议论文集（*SIGIR*），第 39–48 页，2020 年。[6] S. Wang, M. Yu, X. Guo, Z. Wang, T. Moon, B. Chang, and J. Gao, “R^3：用于开放域问答的强化式阅读器-排序器”，载于 第 *325*988, 2018 5981届 *AAAI* 人工智能会议（*AAAI*），第 – 页，.[7] Q. Ai, Y. Chen, Y. Zhang, and W. B. Croft, “用于排序优化的深度列表式上下文模型学习”，载于 第 *42* 届 *ACM* 国际信息检索研究与发展会议论文集（*SIGIR*），第 135–144 页，2019 年。[8] R. Nogueira and K. Cho, “使用 BERT 进行段落重排序”，arXiv:1901.04085，2019 年。[9] L. Xiong, Y. Sun, J. Chen, T. Koo, J. Wang, K. Zhou, V. Pan, and Y. Liu, “面向开放域问答的答案聚焦且位置感知的神经检索器”，载于 北美计算语言学协会分会：人类语言技术会议论文集（*NAACL-HLT*），第 3844–3854 页，2021 年。[10] S. Lu, M. Zhou, D. Shao, Z. Liu, J. Zhang, and H. Chen, “通过强化学习改进多文档摘要的文档选择”，载于 计算语言学协会会议论文集：*ACL-IJCNLP* 专辑，第 3967–3977 页，2021 年。[11] Z. Zhang, Z. Chen, J. Zhao, and H. Chen, “Lite-ranker：用于高效文档排序的轻量级 Transformer 架构”，载于 北美计算语言学协会分会会议论文集（*NAACL*），第 2242–2254 页，2022 年。[12] V. Sze, Y. Chen, T. Yang, and J. Emer, “深度神经网络的高效处理：教程与综述”，*IEEE* 会议录，第 109 卷，第 12 期，第 2463–2492 页，2021 年。[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “近端策略优化算法”，arXiv:1707.06347, 2017 年。[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore：使用 BERT 评估文本生成”，载于 国际学习表征会议（*ICLR*），2020 年。[15] S. Mishra, Y. Liu, D. Brahman, D. Goyal, and S. Ray, “TURQuoISe：领域偏移下问题生成的基准测试”，载于计算语言学协会会议论文集：*EMNLP* 专辑，第 4174 4188 页，2021 – 年。[16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD：用于文本机器阅读理解的 100,000+ 个问题”，载于 自然语言处理实证方法会议论文集（*EMNLP*），第 2383 2392 页，2016 – 年。[17] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “课程学习”，载于第 *26*48, 2009 41届国际机器学习会议（*ICML*），第 – 页。[18] T. Wolf 等人，“Transformers：自然语言处理的最新进展”，载于 自然语言处理实证方法会议论文集：系统演示（*EMNLP*），第 38–45 页，2020 年。