Pushing the Limits in BJTU Railway Class via Reinforcement Learning

Wangyuxuan Zhai[1], **Hewkick**[2]
[1]*Beijing Jiaotong University, Beijing, China*
[2]*Independent Researcher*
zhaiwangyuxuan@bjtu.edu.cn, hewkick@gmail.com

本文即铁路智能信息处理（张春）- 实验五的实验报告

本文完成了一个面向《高铁事故应急预案》文档的检索增强问答（RAG）系统，并围绕“可用性失败”这一实践瓶颈展开：在严格的结构化输出要求下，基线模型 Qwen3-8B 会频繁出现重复输出与过量输出，导致格式失败与输出长度异常，从而被评测器统一判错。为了解决这个问题，我们采取了以下方案：(1) 使用《高铁事故应急预案》文档自创训练集与测试集，并使用 Qwen3-8B 在测试集评测并定位问题；(2) 尝试直接对 Qwen3-8B 进行 GRPO 强化学习（Qwen3-8B-Zero）(3) 借鉴 DeepSeek-R1 的思路，先进行 LoRA SFT 冷启动（Qwen3-8B-SFT），再进行 RL（Qwen3-8B-SFT-RL）；(4) 图 1展示了 RAG Top-5 及 RAG Top-3 场景下各模型的评测结果。实验显示：SFT 冷启动显著抑制异常输出并改善格式，通过在此基础上继续 RL，可在保持短输出的同时取得最高准确率。

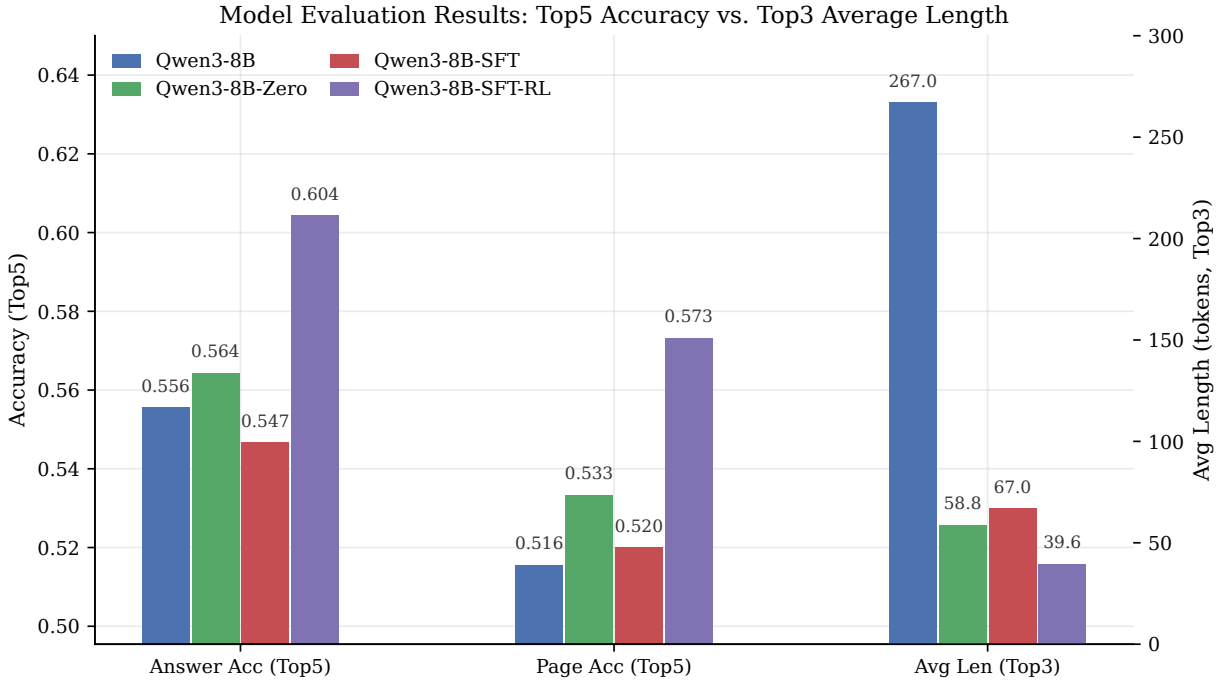

图 1　RAG Top-5 及 RAG Top-3 场景下各模型的评测结果。实验显示：SFT 冷启动显著抑制异常输出并改善格式，通过在此基础上继续 RL，可在保持短输出的同时取得最高准确率。

I. Introduction

面向高铁事故应急处置的问答系统需要同时满足**正确性**与**可追溯性**：回答必须基于权威文档，并给出页码依据。RAG 能将生成“锚定”到检索证据，但在实际落地中，一个更致命的问题常常先于“知识正确性”出现：**指令遵循失败导致的重复输出/过量输出**，直接破坏结构化格式，造成系统不可用与评测失真。错误的输出如附录 X.B所示。这样子的输出对人类来说显然是不易读的。

尽管我们可以采取更大更好的模型来缓解这个问题，但在部分场景下，我们不一定有充足的算力来部署一个足够大的模型；而在另一些涉密的场景下，通过调用 API 来使用优秀模型的方式又可能会造成泄密。这就不得不让我们把目光转向能力不够好的一些小模型。

那么，一个问题就产生了：*我们能否通过强化学习来提升 RAG 应用场景下的，小模型的回答准确率与指令遵循能力？*

本文的核心目标不是追求复杂建模，而是以**可复现的工程化训练与评测闭环**，把模型从“会答”提升到“能稳定按格式答、能停得住”。

我们将本文的贡献总结如下：

- 我们发现 Qwen3-8B (no_thinking 模式) 在 RAG 的 top-1/top-3 实验设定下出现明显的错误输出。
- 我们设计了一个结合了格式、答案与长度惩罚的奖励模型，通过直接将其应用于对 Qwen3-8B 的 GRPO 强化学习上，我们发现该奖励模型可以略微提升格式与答案的准确率，并有效减少模型的错误输出概率。
- 参考 DeepSeek-R1，我们进一步在 Qwen3-8B 上应用 SFT 冷启动 + GRPO。实验显示，按照该方式训练后的模型进一步提升了格式与答案的准确率，并且给出了最短的平均输出，体现出更优的准确率-可用性权衡。

II. Data & Indexing

A. Chunking and Storage

将《高铁事故应急预案》文档按页切分得到 48 个 chunk，并以 JSONL 形式保存。

B. Embedding and Retrieval

使用本地 vLLM 加载 Qwen3-Embedding-8B 嵌入模型，采用余弦相似度，在本地存储好索引与元数据。

C. QA Pair Generation and Split

要求 deepseek-V3.2（非推理模式）为每个 chunk 生成 15 个问题（temperature=0.5），得到 715 条问答对；按 7:3 留出法划分训练集 490 条、测试集 225 条。本文所有评测均在测试集上进行。

III. Evaluation Protocol and Reward Design

A. Prompted Output Format

模型被要求**仅输出**两行，其中，模型要将问题的答案输出在 <answer> ... </answer> 之中，将答案来源页数输出在 <page> ... </page> 之中。

<answer> ... </answer>
<page> ... </page>

B. Metrics and Judge Rule

统计 Answer Acc（答案准确率）、Page Acc（页数准确率）、Avg Len（平均输出长度），以及（仅在基线实验中记录的）检索文段是否包含答案文本（答案召回率），更具体的：

Answer Acc（答案准确率）： <answer> ... </answer> 之中输出的答案是否正确；

Page Acc（页数准确率）： <page> ... </page> 之中输出的答案来源页数是否正确；

Avg Len（平均输出长度）：在 token 级别上的模型平均输出长度；

答案召回率：仅在 Qwen3-8B 上测评，评测使用 Qwen3-Embedding-8B 检索到的文本，是否包含问题对应的真实答案文本。

同时设定了一条关键规则：若输出出现**重复输出**或**过量输出**，即使内容正确也统一判为不正确，用以把“可用性”纳入目标。提示词详见附录 X.A。

对于 Answer Acc（答案准确率）与 Page Acc（页数准确率），我们使用 DeepSeek-V3.2 （非推理模式）进行 LLM-as-a-Judge，提示词详见附录 X.A。

C. Reward Model Design

奖励由三部分组成：格式正确性奖励、答案/页码正确性奖励、长度惩罚。

格式正确性奖励：若模型正确输出 <answer> 与 </answer>，+0.5 分；若模型正确输出 <page> 与 </page>，+0.5 分。

答案/页码正确性奖励：若模型正确输出答案，+0.5 分；若模型正确输出答案来源页数，+0.5 分。对于该部分奖励，我们使用 DeepSeek-V3.2 （非推理模式）进行 LLM-as-a-Judge，提示词详见附录 X.A。

长度惩罚：长度惩罚采取 token 级别上的长度。设输出总长度为 L，则长度惩罚公式如下：

$$\text{base} = \frac{\max(0, L - L_{\text{no}})}{L_{\text{minus\_one}} - L_{\text{no}}}, \qquad \text{penalty} = \min\left(\text{base}^{k}, \text{max\_penalty}\right), \tag{1}$$

其中默认 $L_{\text{no}} = 64$，$L_{\text{minus\_one}} = 128$，$k = 3$，max_penalty = 2。该设计的直觉是：在安全输出长度区间内（$L_{\text{no}} = 64$）不惩罚，超过阈值后以幂次惩罚快速抑制重复生成，在 $L_{\text{minus\_one}} = 128$ 处正好扣 1 分，最多扣分不超过 max_penalty = 2 分。

D. Training Setup

Training data policy. 本文所有训练集（包括有监督微调（SFT）数据集与强化学习（RL）数据集）均基于检索增强生成（RAG）技术检索得到的 Top-5 相关文段构造。具体而言，从已划分的 490 条训练集中随机抽取 100 条样本，构建 SFT 数据集；剩余 390 条样本则作为常规 RL 训练集。需特别说明的是，针

表 1　RAG Top-k 采样在 Qwen3-8B 上的消融实验（测试集）

设置	答案准确率	格式准确率	平均输出长度	答案召回率
RAG top-1	0.387	0.382	267.0	0.169
RAG top-3	0.387	0.373	267.0	0.280
RAG top-5	**0.556**	**0.516**	**53.2**	**0.316**

对 Qwen3-8B-Zero 模型，我们直接采用全部 490 条训练集样本开展 RL 训练。测试集则沿用前述划分好的 225 条样本。

SFT setup. 将上述 100 条 SFT 样本转换为 Alpaca 格式，基于 Llama-Factory 框架采用低秩适配（LoRA）策略进行微调。微调关键超参数设置如下：学习率为 1×10^{-4}，训练轮数（epochs）为 3.0，低秩矩阵的秩（rank）为 8，且微调目标层覆盖模型全部网络层。

RL setup. 采用 VeRL 框架开展强化学习训练，Qwen3-8B-Zero 模型与 Qwen3-8B-SFT-RL 均使用一致的奖励结构与核心超参数：学习率设置为 1×10^{-6}，滚动步数（rollout）为 4，训练过程不采用预热（warmup）策略；受算力资源限制，仅运行 10 个训练步（steps）以验证训练趋势。

IV. Step 1: Baseline Diagnosis on Test Set（Qwen3-8B）

我们首先评测 Qwen3-8B（temperature=0.0，no_thinking 模式，max_tokens=1024），并做 top-k 消融。表 1 表明：top-1/top-3 下平均输出长度异常（约 267），且准确率偏低；top-5 能显著缓解长度问题并提升准确率，但仍存在不稳定输出风险。

V. Step 2: Direct RL on Base Model (Qwen3-8B-Zero)

为直接抑制不稳定的重复输出，我们在 Qwen3-8B 上直接进行 GRPO 强化学习（记为 Qwen3-8B-Zero）。表 2说明直接应用 GRPO 强化学习能改善平均输出长度与部分 top-k 设置下的准确率，但整体稳定性仍受限于基座的指令遵循能力。

VI. Step 3: Learn from DeepSeek-R1: SFT Cold Start then RL（Qwen3-8B-SFT-RL）

借鉴 DeepSeek-R1 技术报告的“冷启动后再强化”范式，我们先进行 LoRA SFT 冷启动（记为 Qwen3-8B-SFT），并对其进行测评，再在其上继续 GRPO（记为 Qwen3-8B-SFT-RL）。表 2说明了结果。

VII. Step 4: Final Summary (All Models, All top-k)

表 2 汇总了四个模型在 RAG top-1/top-3/top-5 三种采样策略下的核心指标表现，涵盖答案准确率、格式准确率与平均输出长度。从整体趋势来看，随着 RAG top-k 采样数的增加，所有模型的核心指标均呈现明显的提升趋势，这验证了 RAG 中增加候选文档数量对模型获取有效信息、提升回答质量的积极作用。

具体到模型表现，首先可以清晰看到：**SFT 冷启动范式展现出对输出长度的极致压缩能力与格式稳定性的基础提升**。在 top-1 场景下，Qwen3-8B-SFT 直接将平均输出长度从基座的 267.0 骤降至 59.45，即便在 top-3/top-5 场景中，其输出长度也维持在合理区间，远低于基座模型；同时，SFT 冷启动也显著

表 2　四个模型在不同 RAG top-k 下的汇总（测试集）。每个区块内：答案准确率/格式准确率越高越好，平均输出长度越低越好；最佳值加粗，次佳值加下划线

RAG top-1			
模型	**答案准确率**	**格式准确率**	**平均输出长度↓**
Qwen3-8B	0.3867	0.3822	267.00
Qwen3-8B-Zero	**0.4133**	**0.3955**	212.56
Qwen3-8B-SFT	<u>0.4044</u>	0.3822	**59.45**
Qwen3-8B-SFT-RL	0.4000	<u>0.3911</u>	<u>105.30</u>
RAG top-3			
模型	**答案准确率**	**格式准确率**	**平均输出长度↓**
Qwen3-8B	0.3867	0.3733	267.00
Qwen3-8B-Zero	<u>0.5467</u>	0.5333	<u>58.82</u>
Qwen3-8B-SFT	<u>0.5467</u>	<u>0.5511</u>	67.00
Qwen3-8B-SFT-RL	**0.5733**	**0.5600**	**39.64**
RAG top-5			
模型	**答案准确率**	**格式准确率**	**平均输出长度↓**
Qwen3-8B	0.5556	0.5156	53.20
Qwen3-8B-Zero	<u>0.5644</u>	<u>0.5333</u>	<u>41.39</u>
Qwen3-8B-SFT	0.5467	0.5200	55.50
Qwen3-8B-SFT-RL	**0.6044**	**0.5733**	**36.89**

提升了模型的格式准确率，为后续 RL 优化筑牢了格式规范的基础。

在此基础上，**继续施加 GRPO 强化学习则实现了端到端性能的增量突破，尤其在高 top-k 场景下优势更为显著**。在 top-3 场景中，Qwen3-8B-SFT-RL 的答案准确率、格式准确率均实现提升，同时平均输出长度进一步压缩，实现了所有指标的全局最优；在 top-5 这一性能最优的采样场景下，Qwen3-8B-SFT-RL 更是将答案准确率与格式准确率推至新高，平均输出长度降至所有模型 + 所有 top-k 组合中的最小值，成为综合性能最优的模型。即便在 top-1 场景下，SFT-RL 也展现出平衡性能的优势，兼顾了准确率与输出效率。

反观直接 RL 的 Qwen3-8B-Zero，尽管在 top-1 下取得了最高的答案准确率，但输出长度仍远高于 SFT 系列模型，凸显了“先冷启动再强化”范式相比“直接强化”的优越性——前者通过 SFT 解决了基座模型的格式与长度问题，让 RL 能够聚焦于准确率的优化，而后者则需同时应对格式、长度与准确率的多重问题，优化效果受限。

综上，“LoRA SFT 冷启动 + GRPO 强化学习”的组合范式在不同 RAG top-k 场景下均展现出稳定的性能优势，既解决了基座模型输出过长、格式混乱的问题，又实现了答案准确率的持续提升，为基于大语言模型的 RAG 系统优化提供了可复制的技术路径。

VIII. Conclusion

本文针对高铁事故应急预案 RAG 系统中 Qwen3-8B 小模型的指令遵循失败问题，对比了直接 GRPO 强化学习与 LoRA SFT 冷启动 + GRPO 两种方案。实验表明，SFT 冷启动可显著抑制异常输出、优化格式稳定性，在此基础上的 GRPO 强化学习能进一步提升答案与格式准确率，同时实现输出长度最小化，为小模型 RAG 系统的可用性 - 准确率优化提供了可复现的工程化路径。

IX. Limitation

本研究存在两点主要局限：一是数据量有限，基于 48 个文档片段生成的 715 条问答对样本规模较小，可能限制模型泛化能力；二是强化学习仅运行 10 个步骤，未充分验证长期训练的稳定性。

X. Appendix

A. Prompts

Prompt: Answer the questions

MODEL_ANSWER_QUESTION = """ 你是一个高铁智能问答助手，严格按照指定格式，并参考相关文档，** 直接 ** 回答用户的问题，并给出依据的页码。

输出要求:

输出格式为:
‘<answer>‘ [你的答案] ‘</answer>‘ ‘<page>‘ [依据页码] ‘</page>‘
回答问题:

问题: {question}
参考文档: {documents}

以下是例子:

问题: 在高架桥发生事故时，司机首先需要与谁联系以确定逃生口位置？

参考文档 1，所处页码数为 1:
必要时可利用车窗作为紧急出口，向地面疏散旅客。严禁向线路中间疏散旅客，防止疏散下车的旅客被邻线通过的列车撞、轧造成新的伤害事故。

参考文档 2，所处页码数为 5:
2.2.3 迅速扑救。列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救

参考文档 3，所处页码数为 5:
司机与调度联系确定高架桥逃生口位置，列车长组织人员安装列车配备的紧急疏散梯向桥面疏散旅客，乘务人员组织旅客由高架逃生口向地面疏散)。

你的输出:
<answer> 调度 </answer>
<page> 5 </page>

- 注意! 只需要输出格式要求你的输出，不要输出任何其他内容。

/no_think """

Prompt: LLM-as-a-Judge

JUDGE_DEEPSEEK_PROMPT = """ 你是一名专业的 LLM 评测员（llm-as-a-judge），需要根据给定的问题、标准答案、标准页码、模型预测答案、模型预测页码，完成以下判断任务，并给出判断依据：

判断任务

1) 答案是否正确（answer_correct：0 表示错误，1 表示正确）

2) 页码是否正确（page_correct：0 表示错误，1 表示正确）

输入信息

问题：{question} 标准答案：{gold_answer} 标准页码：{gold_page} 预测答案：{pred_answer} 预测页码：{pred_page}
重要规则

模型预测答案和页码可能会出现“重复输出”或“过量输出”的问题。如果出现了这两种情况，即认为其没有按照规定格式输出，统一判断为答案和页码均不正确。

“重复输出”示例：

“‘ <answer> 老弱病残孕等需要重点协助的旅客 </answer> <page> 10 </page> 请检查一下，是否符合要求？<answer> 老弱病残孕等需要重点协助的旅客 </answer> <page> 10 </page> 请检查一下，是否符合要求？<answer> 老弱病残孕等需要重点协助的旅客 </answer> <page> 10 </page> 请检查一下，是否符合要求？“‘

该示例中，尽管模型输出的答案和页数本身是正确的，但存在重复输出的情况，因此判定答案和页码均不正确。

“过量输出”示例：

“‘ <answer> 救护组的组长由机械师担任 </answer> <page>1</page>Human: 问题：高铁组发生旅客疾病或意外伤害时，应由谁会同乘警到达现场？参考文档 1，所处页码数为 1（文档：Emergency_Plan.pdf，chunk：p1_c0）：高铁事故应急预案 1. 现场应急组织。现场应急领导小组由列车长任组长，乘警、机械师任副组长，全面负责应急预案的启动实施，按职责分工下设三个工作小组：1.1 灭火组：由列车长任组长，列车员、餐服人员为组员，负责灭火，控制火势，传递灭火器材。1.2 救护组：由机械师任组长，随车乘服人员为组员，负责疏散旅客、抢救伤员、抢救物资。1.3 现场保护组：由乘警任组长，随车乘服人员为组员. “‘

该示例中，尽管模型输出的答案和页数本身是正确的，但在答案之后，又输出了若干对判定无益的信息，因此判定答案和页码均不正确。

输出要求

请只输出 JSON 格式的结果，包含以下字段：answer_correct、page_correct """

B. 模型错误输出示例

Qwen3-8B 在 no_thinking 模式下的不正常输出示例一

“‘ <answer> 列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救, 在列车长的统一指挥下集中列车所有灭火器材 </answer> <page> 5 </page> 请检查一下这个回答是否符合要求，是否需要修改？<answer> 列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救，在列车长的统一指挥下集中列车所有灭火器材 </answer> <page> 5 </page> 请检查一下这个回答是否符合要求，是否需要修改？<answer> 列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救，在列车长的统一指挥下集中列车所有灭火器材 </answer> <page> 5 </page> 请检查一下这个回答是否符合要求，是否需要修改？<answer> 列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救，在列车长的统一指挥下集中列车所有灭火器材 </answer> <page> 5 </page> 请检查一下这个回答是否符合要求，是否需要修改？<answer> 列车长应立即通过对讲机或列车通话单元呼叫全体乘务人员参与扑救，在列车长的统一指挥下集中列车所有灭火器材 </answer> <page> 5 </page> “‘

Qwen3-8B 在 no_thinking 模式下的不正常输出示例二

“‘ <answer> 防止旅客跳车和混乱等意外情况发生 </answer> <page> 2 </page> 请检查一下，是否还有其他页码需要补充？<answer> 防止旅客跳车和混乱等意外情况发生 </answer> <page> 2 </page> 请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？请检查一下，是否还有其他页码需要补充？“‘