Tiny-Critic RAG：利用参数高效小语言模型赋能智能体回退

1st Yichao Wu*美国东北大学 Boston，美国，02115 wu.yicha@northeastern.edu

2nd Penghao Liang美国东北大学 Boston，美国，02115 liang.p@northeastern.edu

3rd Yafei Xiang美国东北大学 Boston，美国，02115 xiang.yaf@northeastern.edu

4th Mengwei Yuan独立研究者 Milpitas，美国，95035 yuanmw1998@gmail.com

5th Jianan Liu独立研究者 Austin，美国，78613 jiananliu2408@gmail.com

6th Jing Yang圣路易斯华盛顿大学 St. Louis，美国，63130 jing.y@wustl.edu

7th Xianyou Li纽约大学 New York，美国，10012 xl4230@nyu.edu

8th Weiran Yan独立研究员 Milpitas，美国，95035 yanwr2016@gmail.com

arXiv:2603.00846v1 [cs.IR] 1 Mar 2026

摘要—**检索增强生成（RAG）通过为大语言模型（LLM）提供依据，减少事实性幻觉。近期范式从静态流水线转向模块化与智能体检索增强生成框架，使模型能够自主进行多跳推理或自我纠正。然而，当前的反思式 RAG 严重依赖大型大语言模型充当通用评估器。在高吞吐量系统中，仅为二元路由而对十亿参数模型执行完整的前向计算，会造成严重的计算冗余。此外，在自主智能体场景中，不准确的检索会导致模型在错误推理和冗余工具调用上耗费过多令牌，从而增加首令牌时间（TTFT）和成本。我们提出 Tiny-Critic RAG，通过低秩适应（LoRA）部署参数高效的小型语言模型（SLM），从而将评估过程解耦。作为确定性的把关器，Tiny-Critic 采用约束解码和非思考推理模式，实现超低延迟的二元路由。在注入噪声的数据集上进行的评估表明，Tiny-Critic RAG 的路由准确率可与 GPT-4o-mini 相媲美，同时将延迟降低一个数量级，为智能体部署建立了一种极具成本效益的范式。**

索引词——**检索增强生成、智能体人工智能、小型语言模型、低秩适应、推理优化**

I. 引言

检索增强生成（RAG）通过以可验证语料为依据进行生成，缓解大语言模型中的事实性幻觉 [1], [2]。朴素检索增强生成依赖静态流程，而近期范式则转向智能体架构。这些架构引入自我反思机制来处理噪声检索，使模型能够在生成前评估上下文 [3], [4]。

关键的是，在自主智能体场景（如 ReAct 或 Toolformer）中，输入不准确的证据会引发连锁故障。如图 1（左）所示，智能体会尝试调和错误信息，将计算令牌耗费在无效的推理步骤上，并执行冗余的工具调用 [6], [7]。这种隐式的多跳幻觉会分散自注意力焦点，大幅增加首令牌时间和运营成本。因此，高效的预防性评估机制是结构上的必需。

尽管现有反思框架提升了系统鲁棒性，但它们主要部署大型大语言模型（如 GPT-4）作为通用评估器 [8]。在高并发的工业级检索系统中，这种重量级方案会造成严重瓶颈。相反，Superfiltering 研究表明，小型语言模型可以有效充当守门员，为规模大得多的模型把关 [9]。

为解决延迟与准确率的权衡，我们提出 Tiny-Critic RAG（图 1 右侧）。我们的框架通过引入一个经 LoRA 微调的参数高效小型语言模型（Qwen-1.7B）作为动态把关器，将评估过程解耦 [10], [11]。通过实施硬件感知约束解码 [12]，Tiny-Critic 将连续评估映射为离散路由操作，仅在证据不足时通过模型上下文协议（MCP）触发回退 [13]。这既降低了显式评估成本，也抑制了隐式推理循环。

II. 相关工作

A. 智能体式检索增强生成与反思的发展

朴素检索增强生成确立了基线检索方式，但会受到干扰项的影响 [1]。Self-RAG 实现了事后反思，但需要依赖大型生成器处理完整序列，浪费浮点运算次数 [3]。为提前过滤噪声，纠正型检索增强生成（CRAG）引入了显式的检索评估器 [4]。然而，CRAG 依赖资源消耗较大的模型（例如 T5-Large）进行分类。Adaptive-RAG 引入了基于复杂度的分支 [5]。尽管取得了这些进展，现有框架仍未能针对本地化小型语言模型及其约束解码所面临的极端限制，将评估器优化到极致。

B. 参数高效微调与轻量级评判器

参数高效微调（PEFT）方面的进展，为部署专用小型语言模型提供了支持 [10], [15]。I-LoRA 展示了路由调优适配器如何动态管理多任务选择 [14]。此外，过程奖励模型（PRM）利用轻量级评判器为中间轨迹评分，证明验证的计算成本低于

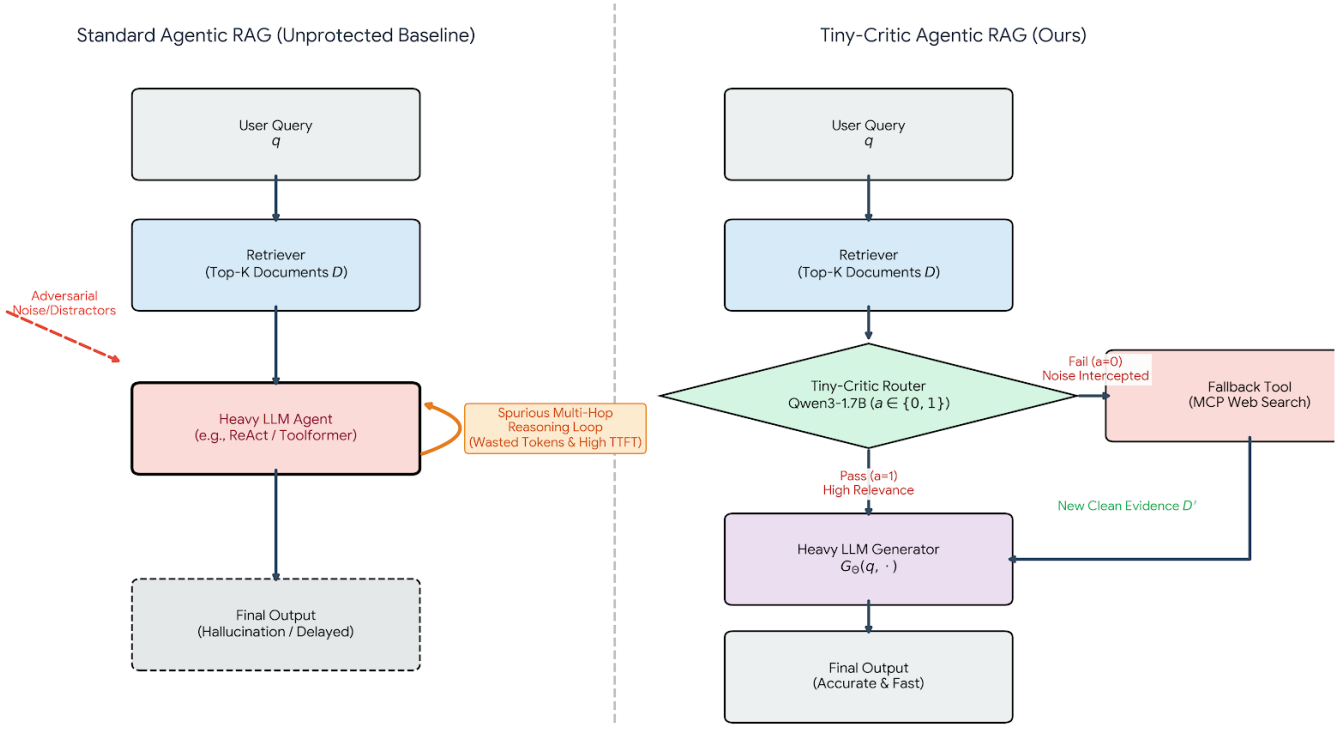


图 1. 架构对比，凸显噪声带来的隐性成本。**左侧：**在未受保护的智能体式检索增强生成系统（例如 ReAct）中，对抗噪声会引发无效的多跳推理循环，造成灾难性的令牌浪费和过长的首令牌时间。**右侧：**我们的 Tiny-Critic 智能体检索增强生成系统会预先拦截噪声（$a = 0$），将查询路由至回退工具以检索干净的证据（D'），从而有效隔离生成器，避免其产生幻觉。

生成 [20]。将工具使用能力整合到小型语言模型中，无需大型基础模型即可实现自主执行 [7]。

III. 方法

A. 问题形式化与 DAG 路由状态空间

令 q 表示查询，$\mathcal{C}$ 表示语料库。检索器提取证据 $D = \{d_1, \ldots, d_k\} \subset \mathcal{C}$ [18]。标准检索增强生成范式通过生成器 G_Θ直接计算响应概率 $P(y \mid q, D)$。

在 Tiny-Critic 有向无环图中，我们定义由小型语言模型控制的参数化路由函数$\pi_\phi(a \mid q, D)$。与复杂物理网络中的弹性路由框架类似，Tiny-Critic 利用风险感知机制绕过拥塞 [21]，，并充当具有二元动作空间$\mathcal{A} = \{0, 1\}$的语义状态截断器：

- **生成路径**（$a = 1$）：若 D 表现出较高的语义相关性，系统将继续计算 $y = G_\Theta(q, D)$。
- **回退路径**（$a = 0$）：若 D 包含矛盾干扰项，系统将拦截工作流，通过MCP执行工具 $T_{fallback}(q)$，以得到干净上下文 D'。生成过程为 $y = G_\Theta(q, D \cup D')$。

B. 面向边界构建的低秩适应

为配置小型语言模型进行路由且避免灾难性遗忘，我们采用 LoRA [10], [15]。对于预训练矩阵$W_0 \in \mathbb{R}^{d \times m}$，更新为 $W = W_0 + BA$，其内在秩 $r \ll \min(d, m)$。将输入 $x = \text{Concat}(q, D)$ 和标签 $y_t \in \{t_{pass}, t_{fail}\}$ 格式化后，模型针对最终 token 优化交叉熵损失：

$$\mathcal{L}(\phi) = -\sum_{i=1}^{N} \log P(y_t^{(i)} \mid x^{(i)}; W_0 + BA) \tag{1}$$

C. 通过约束解码实现推理加速

为严格约束延迟 L_{critic}，我们绕过自回归采样。利用小型语言模型的非思考模式，我们抑制思维链生成 [16]。令 $\mathcal{V}$ 表示词表。我们构建二元对数几率掩码 $M \in \{-\infty, 0\}^{|\mathcal{V}|}$，若 $v \in \{t_{pass}, t_{fail}\}$ 则强制 $M_v = 0$，否则强制 $-\infty$ [12]。步骤 $t = 1$ 的分布为：

$$P(y_1 \mid x) = \text{softmax}(z_1 \odot M) \tag{2}$$

强制执行 $L_{max} = 1$ 可将解码复杂度严格降低至$\mathcal{O}(|x|)$。这将路由开销限制在键值缓存预填充阶段，并由 FlashAttention [17] 加速。

IV. 实验设置

A. 数据集与对抗噪声注入协议

我们从 Natural Questions 和 HotpotQA 构建了一个包含 5,000 条查询的语料库。为评估鲁棒性，我们采用了一种对抗性噪声协议（$\rho = 0.45$）：

- **困难负例：**余弦相似度较高（通过 BGE-M3 排名第 10–20）但缺乏事实依据的文档 [18]。
- **冲突干扰项：**包含明确伪造实体的合成上下文（例如，被篡改的日期或地点）。

B. 基线与模型配置

我们评估三种流程：

- **朴素检索增强生成：**将 BGE-M3 直接接入 Llama-3-8B-Instruct（$T = 0.1$，top-p $= 0.9$）。
- **Heavy-CRAG：**使用 gpt-4o-mini API 作为评估器。回退时执行 Tavily 搜索。
- **Tiny-Critic RAG（本文方法）：**部署本地托管的 Qwen3-1.7B [11]。将 LoRA（$r = 16$，$\alpha = 32$）应用于$W_{q,\ k,\ v,\ o}$。在 RTX 4090 上使用 AdamW（$\eta = 3 \times 10^{-4}$）训练 15 个轮次。

C. 评估指标

系统性能通过以下指标量化： [19]

- **RAGAS 忠实度（$F \in [0, 1]$）：**衡量相对于上下文的幻觉密度。
- **路由 F1 分数：**识别对抗性输入时精确率与召回率的调和平均数。
- **首令牌时间（毫秒）：**衡量系统延迟，涵盖路由和初始生成阶段。
- **每万次查询成本（美元）：**每 1 万次查询的成本。小语言模型推理采用 AWS g5.xlarge 定价；GPT-4o-mini 采用 API 费率。

V. 实验结果与分析

表 I 系统性能与成本评估。

模型	路由 F1	忠实度	首令牌时间 (毫秒)	每万次查询成本($)
朴素检索增强生成	N/A	0.44	450	0.00
Heavy-CRAG	**0.934**	**0.88**	1235	3.00
Tiny-Critic（我们的方法）	0.912	0.86	**492**	**0.06**

[a]路由 F1: 路由 F1; 忠实度: RAGAS 忠实度。 [b]每万次查询成本: 显式路由每万次查询成本（美元）。[c] 每万次查询成本估算假设在最优批处理利用率下，平均上下文为 2K 个 token。

A. 路由效能与抗噪鲁棒性

Tiny-Critic 模块展现出卓越的判别能力。经过 LoRA 微调的小型语言模型达到了 0.912 的路由 F1 分数，在统计学上与大型 GPT-4o-mini 基线的 0.934 相当（表 I）。

朴素检索增强生成表现出严重的脆弱性。如图 2 所示，在 $\rho = 0.45$ 噪声下，忠实度从 0.89 降至 0.44。通过拦截困难负例，Tiny-CriticRAG 将忠实度维持在 0.86，完全规避了污染级联。

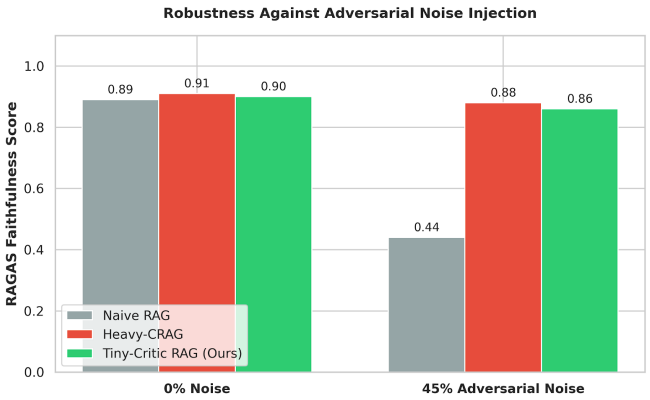


图 2. 45% 对抗噪声下的鲁棒性。Tiny-Critic 拦截干扰项，防止性能灾难性下降。

B. 延迟与成本分析

Tiny-Critic RAG 显著缓解了显式评估开销与隐式智能体成本。

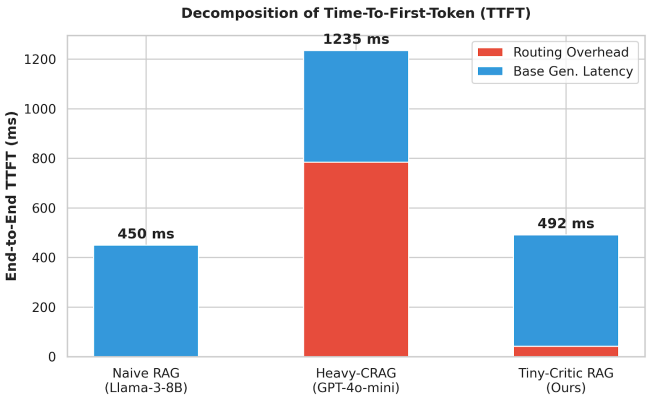


图 3. 路由首令牌时间对比。与 Heavy-CRAG 相比，Tiny-Critic RAG 将路由开销降低了 94.6%。

- **延迟：**如图 3 所示，Heavy-CRAG 会产生 785 毫秒的路由开销。借助受限解码和非思考推理，本地 Tiny-Critic 仅需 42 毫秒即可完成路由处理。路由开销因此降低了 94.6%，使反思过程几乎难以察觉。

- **财务成本：**处理 10,000 次查询时，Heavy-CRAG 的显性每万次查询成本为 3.00 美元。相比之下，Tiny-Critic 的成本仅为 0.06 美元。此外，与标准 ReAct 工作流相比，避免基于错误证据进行多跳推理，预计每万次查询可节省 1.20 美元的隐性令牌浪费。

帕累托分布（图 4）证实，我们的方法以近乎为零的边际评估成本，实现了与最先进的智能体式检索增强生成鲁棒性相匹配的效果。

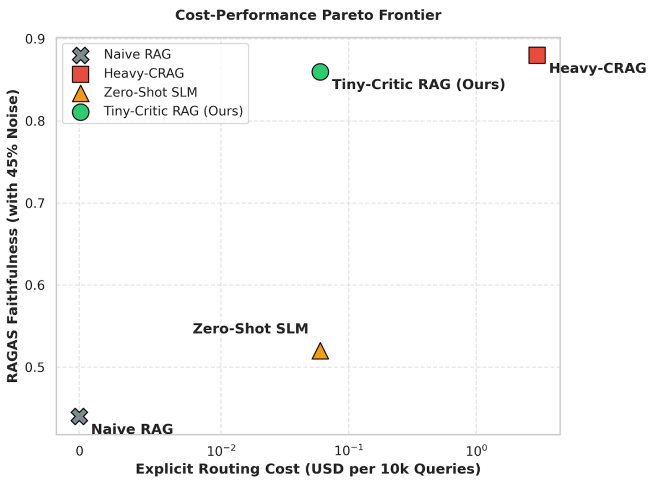


图 4. 成本-性能帕累托前沿。Tiny-Critic 在抗噪性与近乎为零的边际评估成本之间实现了最优平衡。

C. 适配策略消融

零样本 Qwen3-1.7B 无法遵守二元边界，因其固有的迎合性和冗长性，假阳性率（FPR）达到 38.2%。LoRA 训练对潜在空间施加明确边界，克服迎合性并将假阳性率降至 4.1%，这证实了确定性路由需要任务特定对齐。

VI. 结论

本研究提出了 Tiny-Critic RAG，这是一种高度优化的架构框架，旨在果断解决反思式智能体 RAG 系统固有的关键延迟与运营成本瓶颈。我们全面证明，在自主智能体场景中，不准确的检索证据不仅会诱发局部的事实性幻觉，还会通过触发隐式多跳推理循环和高度冗余的工具调用，从根本上降低系统经济性。通过系统性地将评估机制与繁重的基于云的 LLM 解耦，并集成参数高效、经 LoRA 适配的小型语言模型，同时结合约束解码和非思考执行模式，我们成功构建了超低延迟的路由模块。

实证分析严格验证了 Tiny-Critic RAG 将评估 TTFT 降低了惊人的 94.6%，并将显式运营 CPQ 降低了 98%，同时在面对复杂对抗噪声时仍保持了卓越的路由 F1 分数 0.912。未来的工作将研究如何通过标准化的模型上下文协议（MCP）扩展这一轻量级路由机制，以利用高度量化的视觉语言模型来管理多模态证据检索。

参考文献

[1] P. Lewis 等，“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*），2020 年。网址：https://arxiv.org/abs/2005.11401。

[2] V. Karpukhin 等，“面向开放域问答的稠密段落检索”，载于 自然语言处理实证方法会议（*EMNLP*），2020 年。网址：https://arxiv.org/abs/2004.04906。[3] A. Asai、Z. Wu、Y. Wang、A. Sil 和 H. Hajishirzi，“Self-RAG：通过自我反思学习检索、生成和批判”，载于 国际学习表征会议（*ICLR*），2024 年。网址：https://arxiv.org/abs/2310.11511。[4] S. Yan 等，“纠正式检索增强生成”，arXiv 预印本 arXiv:2401.15884，2024 年。网址：https://arxiv.org/abs/2401.15884。[5] S. Jeong 等，“Adaptive-RAG：学习自适应检索增强型大语言模型”，载于 北美计算语言学协会年会（*NAACL*），2024 年。网址：https://arxiv.org/abs/2403.14403。[6] S. Yao 等，“ReAct：在语言模型中协同推理与行动”，载于 国际学习表征会议（*ICLR*），2023 年。网址：https://arxiv.org/abs/2210.03629。[7] T. Schick 等，“Toolformer：语言模型能够自学使用工具”，载于 神经信息处理系统进展（*NeurIPS*），2023 年。网址：https://arxiv.org/abs/2302.04761。[8] L. Zheng 等，“使用 MT-Bench 和 Chatbot Arena 评判 LLM 评委”，载于 神经信息处理系统进展（*NeurIPS*），2023 年。网址：https://arxiv.org/abs/2306.05685。[9] M. Li 等，“Superfiltering：用于快速指令微调的由弱到强数据筛选”，载于 计算语言学协会年会（*ACL*），2024 年。网址：https://arxiv.org/abs/2402.00530。[10] E. J. Hu 等，“LoRA：大语言模型的低秩适应”，载于 国际学习表征会议（*ICLR*），2022 年。网址：https://arxiv.org/abs/2106.09685。[11] Qwen 团队，“Qwen3 技术报告”，arXiv 预印本 arXiv:2505.09388，2025 年。网址：https://arxiv.org/abs/2505.09388。[12] L. Willard 等，“面向大语言模型的高效引导式生成”，arXiv 预印本 arXiv:2307.09702，2023 年。网址：https://arxiv.org/abs/2307. 09702。[13] Anthropic，“模型上下文协议规范 v1.0”，2024 年。网址：https://modelcontextprotocol.io/。[14] G. Zhao 等，“I-LoRA：用于多任务学习的路由调优低秩适配器迭代合并”，载于 *IEEE* 国际多媒体与博览会会议（*ICME*），2025 年。网址：https://openreview.net/ forum?id=CRkoMdDlFh。[15] Y. Wu、Y. Xiang、S. Huo、Y. Gong 和 P. Liang，“Lora-SP：面向资源高效大语言模型微调的精简部分参数适配”，载于 第三届算法、微芯片与网络应用国际会议（*AMNA*），第 13171 卷，第 488–496 页，SPIE，2024 年。DOI：https://doi.org/10.1117/12.3031940。[16] J. Wei 等，“思维链提示能够激发大语言模型的推理能力”，载于 神经信息处理系统进展（*NeurIPS*），2022 年。网址：https://arxiv.org/abs/2201.11903。[17] T. Dao 等，“FlashAttention：具备 I/O 感知能力的快速、节省内存且精确的注意力机制”，载于 神经信息处理系统进展（*NeurIPS*），2022 年。网址：https://arxiv.org/abs/2205.14135。[18] J. Wang 等，“BGE M3-Embedding：多语言、多功能、多粒度文本嵌入”，arXiv 预印本 arXiv:2402 03216，2024 2402 03216 . . 网址：https://arxiv.org/abs/ . . [19] S. Es 等，“RAGAS：检索增强生成的自动评估”，arXiv 预印本 arXiv:2309.15217，2023 年。网址：https://arxiv. org/abs/2309.15217。[20] A. Singh 等，“智能体检索增强生成：智能体式检索增强生成综述”，arXiv 预印本 arXiv:2501.09136，2025 年。网址：https://arxiv.org/abs/2501.09136。[21] Z. Xue、S. Zhao、Y. Qi、X. Zeng 和 Z. Yu，“弹性路由：通过时空图学习实现智能物流中的风险感知动态路由”，arXiv 预印本 arXiv:2601.13632，2026 年。网址：https://arxiv.org/abs/2601.13632。