基于检索增强生成的 Sigma 规则到 MITRE ATT&CK 的自动映射方法

陈烨楷[1]，田玉丹[2]，赵明昊[1]，钱卫宁[1]
1. 华东师范大学 数据科学与工程学院，上海 200062；
2. 华东师范大学 信息化治理办公室，上海 200062

摘要

Sigma 规则是描述日志检测逻辑的通用格式，MITRE ATT&CK 框架则以“战术–技术”结构归纳对手行为，实现两者的自动映射对于威胁检测具有重要意义。然而，Sigma 规则语义分散，且常覆盖多个技术标签，现有方法难以兼顾效率与准确性。针对这一问题，大语言模型为此类跨域语义映射提供了新的技术路径，但通用模型受限于领域知识储备，仍难以可靠地区分细粒度技术标签。为此，提出了 RAG–S2A 框架，通过 BM25 与稠密检索融合召回候选，结合多正样本对比学习与困难负样本挖掘实现领域语义对齐，将检索准确率从 20.26% 提升至 72.75%。相较于 F1 为 26.69% 的 BERT 多标签分类基线，该方法在 Qwen3–4B 上取得 60.59% 的 F1。结果表明，检索增强与领域适配的结合能够有效提升资源受限模型在该任务上的映射精度。

关键词

网络安全；大语言模型；检索增强生成；Sigma 规则；MITRE ATT&CK；威胁检测

中图分类号：TP391　　文献标志码：A　　doi：10.11959/j.issn.2096–0271.xxxxxxx

An automated retrieval-augmented generation method for mapping Sigma rules to MITRE ATT&CK

CHEN Yekai[1], TIAN Yudan[2], ZHAO Minghao[1], QIAN Weining[1]
1. School of Data Science and Engineering, East China Normal University, Shanghai 200062, China;
2. Informatization Governance Office, East China Normal University, Shanghai 200062, China

Abstract

Sigma rules provide a generic format for describing log detection logic, whereas the MITRE ATT&CK framework organizes adversary behaviors in a tactics-techniques hierarchy. Automating the mapping between the two is important for threat detection. However, the semantics of Sigma rules are dispersed, and a single rule often covers multiple technique labels, making it difficult for existing methods to balance efficiency and accuracy. To address this issue, large language models provide a new technical path for cross-domain semantic mapping, yet general-purpose models still

struggle to reliably distinguish fine-grained technique labels because of limited domain knowledge. To this end, the RAG-S2A framework was proposed. BM25 and dense retrieval were combined for candidate recall, and multi-positive contrastive learning with hard negative mining was employed for domain semantic alignment, improving retrieval accuracy from 20.26% to 72.75%. Compared with the BERT multi-label classification baseline (26.69% F1), the method achieved an F1 score of 60.59% on Qwen3-4B. The results indicate that the integration of retrieval augmentation and domain adaptation can effectively improve mapping accuracy for resource-constrained models.

Key words
cybersecurity, large language model, retrieval-augmented generation, Sigma rule, MITRE ATT&CK, threat detection

0 引言

随着网络攻击的频率与复杂度不断上升，安全运营中心（security operations center，SOC）必须对威胁行为做出迅速、准确的响应。Sigma 规则采用 YAML 格式描述日志来源、匹配条件和告警逻辑，可被转译至各类安全信息和事件管理系统（security information and event management，SIEM）或检测平台中，从而降低跨产品迁移的成本。同时，MITRE ATT&CK 以战术、技术或子技术结构归纳对手行为，为威胁狩猎、检测覆盖度评估、溯源分析等提供了统一的语义坐标。将 Sigma 规则自动映射到 ATT&CK 标签上，能够建立起对手行为和检测工程之间的可量化映射关系，对威胁检测、攻击归因和防御策略制定具有实际意义[1-2]。

然而，该映射任务仍面临许多难题。人工标注的工作量大、耗时长，各个标注者主观差异较大，且一致性不高，不能应对海量规则的标注工作。CardinalOps 2025 年发布的 State of SIEM Detection Risk 报告显示，企业摄取到的日志遥测理论上可以支撑约 90% 的 ATT&CK 技术检测，在规则工程方面只有 21% 的技术获得了检测规则覆盖，并且还有一定比例的失效规则，说明可观测性与可检测性之间存在着系统性的差距[3]。

另一方面，目前的自动化方法对于语义的理解还比较欠缺。基于关键词匹配或规则的方法无法捕捉到深层次的语义联系，准确率较低；传统的机器学习方法（如使用 BERT [4]的分类器等）大多面向非结构化文本进行设计，不能直接运用于 Sigma 规则中的结构化字段信息；尽管大语言模型（large language model，LLM）具备较强的通用语义理解能力，但在零样本或少样本场景下，但通用模型受限于其领域知识储备不足，其对 ATT&CK 框架中细粒度技术的区分能力依然有限。

这一挑战的根源在于 MITRE ATT&CK 标签空间本身具有大规模、高细粒度及强语义重叠的特征。2025 年 4 月发布的 ATT&CK v17 包含 14 个战术，211 个技术及 468 个子技术，大部分技术和子技术之间存在语义重叠或层级关系。例如 T1053.005 （Scheduled Task） 与 T1569.002（Service Execution）虽然都涉及系统级任务执行，但适用场景不同。准确分辨这些细微差别，需要对技术的定义和适用场景有充分的理解。在有限的资源条件下，如何选择合适的模型规模以及

用技术手段来弥补模型的不足，是实际部署中需要面对的问题[5]。

为应对上述挑战，本文提出RAG-S2A自动映射框架，整体流程包括预处理与查询构建、混合检索、重排序、受控生成与后处理四个模块。其核心思路是在BM25稀疏检索与经对比学习微调的稠密检索基础上构建高质量上下文，并结合重排序与生成约束机制提升细粒度技术映射的稳定性与准确性。在此基础上，本文的主要贡献体现在三个方面：1）提出面向Sigma规则到ATT&CK映射任务的RAG-S2A框架，通过混合检索与领域适配的检索模型为大语言模型提供高质量上下文，有效缓解纯生成方法的幻觉问题；2）提出基于对比学习与困难负样本挖掘的双编码器领域适配方法，将检索准确率从19.45%提升至75.25%，并构建开源可复现的实验基准；3）对不同规模模型及RAG感知微调策略的性能差异进行系统评估，发现RAG对中小模型增益显著（F1相对提升超96%），为资源受限场景下的模型选择提供实证依据[6-8]。

1 背景与预备知识

1.1 MITRE ATT&CK框架

MITRE ATT&CK是通过对真实攻击活动的分析来形成对手行为知识的体系，涉及企业环境、移动端以及工业控制等各个领域[1]。该框架用分层的方式组织攻击行为，战术（Tactics）描述攻击者的目标，如初始访问、权限提升等，技术（Techniques）描述达成战术目的的方法，子技术（Sub-techniques）给出更加精细的实现细节。每个技术条目中包含着大量的描述信息，如检测方法、缓解措施以及数据源建议等，这些信息作为检测规则语义映射的关键依据[1,5]。

1.2 Sigma检测规则

Sigma是一种平台无关的检测规则格式，用YAML结构来描述日志事件中的可疑行为模式[2]。与SIEM平台所特有的规则语言不同，Sigma规则可以转换为多种安全产品的规则。规则由logsource（定义日志类型与平台）和detection（定义检测逻辑）两部分组成，支持使用tags字段来标注对应的ATT&CK技术标识符。然而在实际场景中，tags字段常常为空或标注有误，这是本研究所要解决的核心问题[2]。

1.3 检索增强生成

检索增强生成（retrieval-augmented generation，RAG）是一种把信息检索和生成式模型结合起来的混合范式。其核心思路是在生成之前，先从外部知识库中检索相关文档来增强上下文，从而缓解纯生成模型的幻觉问题。典型的RAG框架包含检索器和生成器两个核心模块：检索器负责从语料库中召回与查询相关的文档片段，生成器则基于检索到的上下文进行推理并产出答案。检索器的实现既可以包括基于词频统计的稀疏检索方法（如BM25[7]），也包括基于预训练语言模型的稠密检索方法[8-10]，后者通过将文本映射到连续向量空间来捕捉语义相似性。

1.4 双编码器与对比学习

双编码器（bi-encoder）属于常用的语义检索建模方式[8]，分别使用查询编码器与文档编码器对查询和候选文档进行独立编码，将它们映射到同一个向量空间中，

然后利用向量相似度来进行相关性度量和检索。与交叉编码器相比，双编码器支持离线编码候选文档，并能结合近似最近邻索引实现大规模检索，因此效率和可扩展性更好。

在训练目标上，对比学习常用于改善双编码器的表示空间。通过构造查询正样本文档匹配对，以及若干查询负样本文档不匹配对，使得模型倾向于提高匹配对的相似度，降低不匹配对的相似度。促使模型习得更具有判别力的向量表示，使检索阶段可以更好地分辨出语义相近但标签不同的候选项[8,11-12]。

2 相关工作

尽管业界和学术界已经尝试将各种安全文本或者检测规则映射到 MITRE ATT&CK，但是针对 Sigma 规则这种结构化输入的公开可复现研究还比较少见。综合已有工作可以归纳为四类，分别是①维护人工标注的知识库，②表征学习驱动的检索与分类，③大语言模型与检索增强生成，④工具链与商业系统。由于任务场景、输入类型、标注粒度（战术、技术、子技术）存在差异，难以横向比较，但是各方法的假设与局限具有可比性。

2.1 手工标注与静态映射

安全社区早期对 Sigma 规则的映射实践主要依赖人工维护的静态映射表，而非自动化的规则模板匹配。例如，MITRE CTID 的 TRAM[13]平台可以自动映射威胁情报报告到 ATT&CK，但是其开箱即用的模型只包含大约 50 个常用 ATT&CK 技术条目，需要针对特定领域手工标注扩展数据才能定制。Sigma 社区维护的 sigma_attack_nav_coverage 等文件也尝试建立规则和技术的静态联系。这类方法依靠领域专家的经验，泛化能力差，当规则的描述发生改变或者逻辑比较复杂的时候，很容易漏标或者误标，而且随着规则数量的增多，维护成本会急剧上升[14]。潘亚峰等[15]利用自然语言处理技术从 ATT&CK 技术定义文本中提取语义规则，在一定程度上实现了审计日志到攻击技术的自动映射，但该方法仍依赖手工确认的规则模板，且面向系统审计日志而非 Sigma 检测规则。

2.2 基于监督学习的文本分类与检索方法

Husari 等人较早提出 TTPDrill[16]，旨在从非结构化威胁情报（cyber threat intelligence，CTI）文本中自动抽取威胁行为并映射到 ATT&CK。在 CTI 和安全文本分析领域，研究者通常将 ATT&CK 标签预测当作文本分类或者检索问题来研究[17-19]。Alves 等人用 MITRE 公开示例句子微调各种 BERT 模型，预测非结构化文本的技术标签；其中最好的模型在两个数据集上的准确率分别达到了约 82.64%、78.75%。刘晨静等人提出的注意力 Transformer Hierarchical RNN（ATHRNN）使用 Transformer 嵌入和注意力递归结构相结合的方式来建模战术和技术标签，从威胁情报中抽取战术、技术和程序（tactics, techniques, and procedures，TTP），在自建数据集上实现了比基线模型的宏/微 F-score 分别提高 6.5%、8.2% 的成果。Alam 等人提出 LADDER 框架，从 CTI 报告中抽取攻击模式，然后利用本体把它们映射到 MITRE ATT&CK 中。Mărmureanu 和 Oprişa[20]建

立了一个机器学习分类器，把5 000多条结构化查询映射到战术标签上。Rani等人[21]用BERT嵌入和线性分类器从高级持续威胁（advanced persistent threat，APT）报告中抽取TTP，构建出8 387个句子以及50篇文档构成的数据集，句子级F1分数为88%，文档向量级F1分数为75%；后续的TTPXHunter[22]在扩展数据集和模型后性能提升至92%以上。

综上所述，由于Sigma规则包含YAML结构和检测逻辑字段等特有信息，以上方法主要处理的是自由文本或其他类型的安全内容，难以直接映射到Sigma规则上[20-22]。

2.3 基于大语言模型与检索增强的生成式方法

近几年来，大语言模型被广泛地应用于ATT&CK相关任务上，例如零样本/少样本标签推断、知识补全和检索增强生成等。Daniel等人[23]使用ChatGPT、Claude、Gemini等LLM为973条Snort NIDS规则标注ATT&CK战术和技术，结果表明生成式模型可以给出解释性的初始映射，但是纯LLM方法在精度上劣于传统的机器学习方法。Fayyazi等人[24]比较了微调编码器与结合检索增强的解码器在TTP分析中的表现，指出外部上下文有助于提高召回，但精度控制仍是难点。Wudali等人[25]面向结构化SIEM规则构建提示链，并结合外部知识检索与多模型推荐进行标签标注，表明检索证据对安全规则映射任务具有积极作用。相比之下，AttacKG+[26]和TechniqueRAG[27]更偏向安全知识组织或通用安全检索增强，并非直接面向Sigma规则到ATT&CK映射任务，刘天扬等[28]将图检索增强生成与少样本学习结合用于美术作品鉴赏，说明RAG已具备跨领域迁移潜力。

从方法演进的角度看，RAG在安全语义理解任务中的作用已由“为生成补充外部知识”逐步转向“提升证据选择、利用与纠错的可靠性”。早期研究主要关注通过检索为模型提供相关上下文，而近期工作则更加重视证据使用过程本身，包括生成阶段的自反思与证据判别、训练阶段对检索与生成的协同建模、对带噪检索上下文的鲁棒适应，以及检索后的纠错与可靠性约束。相应地，Self-RAG[29]、RA-DIT[30]、RAFT[31]和CRAG[32]分别代表了自反思检索、检索-生成协同调优、带噪上下文适应与检索纠错等几类典型优化思路；更近期的SafeRAG[33]与RAG+[34]则进一步从证据可靠性控制、上下文噪声抑制及生成稳健性等方面对RAG进行了扩展与完善。尽管上述方法并非专为Sigma规则到ATT&CK映射任务提出，但其核心目标与本文关注的目标高度一致，即如何提升检索证据质量并增强生成可靠性，因此可为本文的方法设计与实验讨论提供参考。

2.4 工程化工具与商业系统

在工程实践中，一些厂商以及社区工具提供了自动预测ATT&CK标签的功能。以Uncoder AI[35]为例，它在检测工程IDE里具备了一键式预测Sigma规则ATT&CK标签的功能，并且宣称其模型训练所使用的人工标注Sigma规则数目超过20 000条。由于相关模型权重、训练数据和评测基准都没有完全公开，外部研究者不能复现实验结果进行公平比较。另一方面，SigmaGen[36]提出将安全博客或者报告自动生成为Sigma规则并映射到ATT&CK的流程，该工作发现通用LLM直接生成的规则存在语法及逻辑错误，需要结合领域

工具（TRAM）和规则校验来修正和映射。就工程落地而言，这类系统具有丰富的经验，但是大多采用闭源服务、私有评测或者专用平台集成的方式，并不能直接被当做开源可复现的基线使用。

2.5 研究动机与本文贡献

由于各个方法的任务场景、输入类型和数据集不同，不能直接进行量化对比，表1展示了本文方法与相关工作在特点上的对比。

表1 相关工作特点对比

类别	代表方法与系统	输入	标注与训练需求	主要特点	局限性
规则匹配	TRAM[13]、sigma_coverage脚本[14]	Sigma规则（含tags）	无需训练；依赖手工标签	实现简单，成本低	覆盖范围受限；缺少tags时无法推断；难以区分细粒度技术
文本分类	BERT分类器[17]、ATHRNN[18]、TTPHunter[21]	CTI报告/安全文本	需要标注语料/监督训练	适合非结构化文本的TTP抽取	与结构化检测规则差异大；迁移到Sigma需重新建模
LLM映射	Daniel等LLM标注[23]、RAM[25]	Snort/SIEM规则	通常无需训练；依赖提示工程与外部知识	可解释、易迭代，适合快速原型	推理成本高；对模型规模敏感；易受相似技术混淆
混合RAG	TechniqueRAG[27]、RAG式TTP分析[24]	安全文本/知识库	少量标注或无需训练；依赖检索语料质量	检索证据可提升召回与可追溯性	证据噪声会传递到生成；缺少规则结构信号
工程化系统	Uncoder AI[35]、SigmaGen[36]与TRAM	检测规则/安全内容	多为私有数据与闭源评测	集成度高，便于生产使用	可复现性弱；评测不透明；难以作为公开基线
本文方法	RAG-S2A	Sigma规则（移除tags）与ATT&CK描述库	少量标注数据微调检索器；LLM推理可选	混合检索与领域适配提升候选质量；对中小模型增益显著	仍受检索召回上限与标签歧义影响

就目前研究而言，大多集中在CTI报告、自然语言文本的TTP识别上，或者使用私有模型和闭源系统，同时开源社区缺少可以复现的基准系统和数据集，也影响了该领域的研究。此外，Sigma规则映射也要考虑到结构化字段、专业术语、细粒度技术之间的差异。本文提出的RAG-S2A在开源中小模型上利用检索增强和领域适应的方法得到高性价比的映射效果，为构建可复现基准并推动实际应用提供了一条可行路径。

3 方法

本章主要对RAG-S2A（Retrieval-Augmented Generation for Sigma-to-ATT&CK）方法的设计和实现进行阐述。该任务存在三个方面的主要困难。第一，标签空间规模大且存在语义重叠，许多技术之间有层级关系或者功能相似性；第二，Sigma规则用YAML方式组织，主要由结构化字段定义检测逻辑，但是也伴随少量自然语言元数据，而ATT&CK技术是以非结构化的文本叙述对手行为和适用的情境为主，两者在表示形式以及信息密度上存在着本质的区别；第三，标注数据稀缺，

高质量的规则-技术对标注需要网络安全专家参与，成本高且很难大规模获取。

因此本文采用基于检索增强生成的两阶段框架，检索阶段从ATT&CK知识库中找相关技术描述，生成阶段用检索到的知识来预测标签。相比于直接使用大语言模型的零样本方法，该范式通过引入外部权威知识库来缓解模型幻觉的问题；相较于传统的监督学习，该范式不需要大量的标注数据就可以适配新出现的技术标签[6]。

3.1 系统架构

如图1所示，RAG-S2A系统采用流水线式设计，推理阶段主要包含四个顺序执行的模块。① 预处理与查询构建：从原始Sigma规则中抽取并融合关键字段，构建适用于检索的查询表示；② 检索：在ATT&CK知识库中检索候选技术集合。系统支持三种基础检索方式——BM25关键词检索[7]、稠密检索[8]与混合检索。此外，还提供自一致性检索作为增强策略，在混合检索基础上进行多次采样与加权汇总，以提升候选集合的稳定性；③ 重排序（可选）：对候选列表进行精排，使用LLM对相关性进行推理式评分，从而过滤噪声并改善Top-K排序质量；④ 受控生成与后处理：在检索上下文约束下生成ATT&CK标签集合，并对输出进行标准化、去重与一致性校验，得到最终结果。

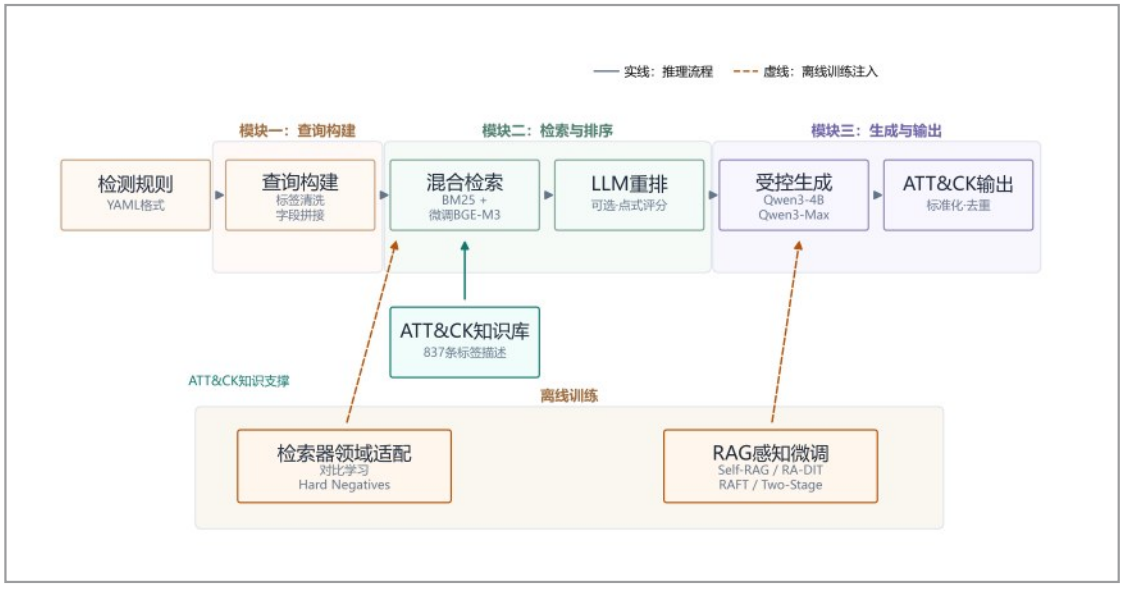


图1 RAG-S2A系统框架

该架构按模块化原则设计，各个部分可以单独优化或替换。主要的设计考量包括：混合检索策略平衡精确匹配与语义理解；领域适应的双编码器提高检索准确率；思维链重排序改善候选排序质量；自一致性检索加强大模型的输出稳定性。以下各节将分别阐述各模块的技术细节。

3.2 预处理与查询构建

输入的Sigma规则使用YAML格式定义检测逻辑，典型字段有logsource表示日志来源（如process_creation、file_change等），detection定义检测条件（selection、filter），title提供规则标题，description描述检测意图。原始规则中常含有tags字段标注过已有的ATT&CK标签，在训练阶段和测试阶段都需要移除该字段，模拟出真实的应用场景下标签预测的任务。

预处理模块包含三个处理步骤，分别为标签移除、字段选择和查询构建。第一步，从规则中删除含有ATT&CK标签的tags等标注信息相关的字段，防止因为泄露标记信息导致模型性能评估偏高。第二步，基于语义贡献度选取语义核心字段，通过字段选择函数从规则所有字段中选出语义核心字段，即title字段包含规则的核心语义、description字段给出检测意图说明、logsource字段限定应用场景、detection字段给出具体的检测逻辑。第三步，查询函数 会将选定字段变成一个统一的查询字符串表示：x_{raw} P_{build}

$$q = P_{\text{build}}(x_{\text{raw}}) = \text{Concat(title, description, logsource, detection)}$$

该设计在保持规则结构化信息的基础上，生成了可被检索模型识别的自然语言表示形式，从而能够涵盖从战术层面到技术层面的所有ATT&CK条目。

3.3 混合检索模块

混合检索模块是RAG-S2A的核心组件，它的性能好坏直接影响到最终映射的准确性。单一检索范式存在固有的局限性，稀疏检索（如BM25等）用精确的术语匹配获得高召回率，能较好地处理专业术语、缩写，如regsvr32、powershell等，但不能捕捉语义等价表述；基于预训练编码器的稠密检索用向量相似度匹配语义相关的文档，但是忽略关键术语的精确匹配信号，在垂直领域的表现不佳。

为兼顾两种范式的优势，本文提出混合检索方法。设ATT&CK知识库为 $Y_{\text{kb}} = \{d_1, d_2, \ldots, d_M\}$，技术数量为。其中M为技术条目总数。给定查询q，检索函数定义为：

$$R(q) = \text{TopK}\left(\left\{\alpha s_{\text{dense}}(q,d) + (1-\alpha) s_{\text{sparse}}(q,d) : d \in Y_{\text{kb}}\right\}\right)$$

s_{dense} 为双编码器产生的余弦相似度，s_{sparse} 为BM25分数，α 为融合权重。该权重通过网格搜索确定，搜索范围为[0.5, 0.9]，步长0.1。在验证集上最优配置为 $\alpha = 0.8$，表明稠密检索承担主要作用，而稀疏检索提供必要的精确匹配信号。

3.3.1 稀疏检索：BM25

BM25算法基于词频-逆文档频率（term frequency-inverse document frequency，TF-IDF）框架评估查询和文档之间的匹配度。给定查询和文档，BM25分数计算如下：

$$\text{BM25}(q,d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t,d) \cdot (k_1 + 1)}{f(t,d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}$$

其中f(w,d)为词w在文档d中的词频，|d|为文档长度，avgdl为语料库平均文档长度，k_1 和b为自由参数。本文采用标准配置 k_1=1.5、b=0.75。BM25的优势在于能够精确匹配专业术语，如对“regedit”、“svchost”等安全领域特有词汇能够准确定位相关技术。

3.3.2 稠密检索:领域适应的双编码器

通用嵌入模型在大规模的通用语料上预训练，有很强的自然语言理解能力，但是对于网络安全垂直领域的语义理解能力较弱。经过实验得知，原始BGE-M3对于ATT&CK技术检索任务的准确率只有19.45%，无法满足需求。本文采用对比学习的方式，将微调的BGE-M3模型适配Sigma-ATT&CK映射任务。双编码器架构包含独立参数的查询编码器和文档编码器，分别将规则查询和技术文档编码成向量表示[38]。

$$v_q = \mathrm{E}_q(q), v_d = \mathrm{E}_d(d)$$

相似度度量采用余弦相似度：

$$s_{\text{dense}}(q,d) = \text{cosine}(v_q, v_d)$$

3.3.3 对比学习与困难负样本挖掘

$D_{\text{train}} = \{(q_i, d_i^+)\}_{i=1}^N q_i i d_i^+$ 为了提高规则到ATT&CK标签的语义匹配能力，本文用对比学习对双编码器进行领域适配。训练集是由规则-标签对组成的，记为，其中q 为第i条Sigma规则的文本表示，d + 为其对应的ATT&CK技术描述（正样本）。模型学习将规则与其正确技术描述映射到相近的表示空间位置，同时与错误标签保持足够的间隔，从而提升检索阶段的可分性[8,11-12]。

由于对比学习效果很大程度上取决于负样本质量，本文综合使用了三种负样本来达到兼顾覆盖性和难度的目的，首先在batch内使用其他样本的正标签作为In-batch负样本来获得大量的对比信号，然后从标签空间随机采样Random负样本来增强负样本的多样性、缓解batch分布偏差，最后使用Hard负样本专门针对“语义相近但标签不同”的样本进行训练，使模型在细粒度技术概念上具备更强的判别能力。

其中，困难负样本是离线方式预生成的。具体做法是用预训练的BGE-M3编码所有的ATT&CK技术描述，计算每一个正标签和其他标签之间的余弦相似度；对于每个正标签，从与其相似度最高的候选中选取Top-K个，排除同一规则下的其他正标签后，作为该标签的困难负样本。该策略使模型能够区分语义相近的技术，从而学习到更加稳定的分类边界[39]，例如T1112（Modify Registry）与T1547.001（Registry Run Keys）虽然都涉及注册表操作但属于不同战术。

另一方面，Sigma规则通常对应平均约2.4个的ATT&CK标签，如果仍然用单正样本InfoNCE进行训练，那么其余真实标签就会被当作负样本，产生冲突信号。因此，本文将InfoNCE扩展为多正样本的形式。对于查询q 及其正标签集合{d 1, …, d }，损失函数定义为：

$$L = -\frac{1}{m}\sum_{j=1}^{m}\log\left(\frac{\exp(\text{sim}(q_i, d_i^j)/\tau)}{\exp(\text{sim}(q_i, d_i^j)/\tau) + \sum_{d\in D_{\text{neg}}}\exp(\text{sim}(q_i, d)/\tau)}\right)$$

其中sim(·)表示余弦相似度，τ为温度超参数（本文设为0.05），负样本集合为$D_{\text{neg}} = D_{\text{in_batch}} \cup D_{\text{hard}}$，包含批次内负样本与困难负样本。该目标函数在提升查询与全部真实标签相似度的同时，抑制其与负标签的相似度，从而更契合多标签映射场景下的检索训练需求。

3.4 自一致性检索

为进一步提升预测稳定性，本文引入自一致性检索策略。该策略通过多次独立检索并汇总结果来减少单次检索的偶然性：对同一查询进行 N 次重复检索，每次采用不同的稀疏-稠密融合权重组合以增加候选多样性，最终通过加权分数累积选择得分最高的候选技术作为检索结果。形式化地，给定查询 q，自一致性检索函数定义为：

$$\text{SelfConsistency}(q)=\text{ScoreAgg}(R(q,\alpha_i),i=1,\ldots,N)$$

其中，N 为采样次数，α 为第 i 次检索的融合权重组合，从偏向稠密检索的 α=0.9 均匀过渡到偏向稀疏检索的 α=0.1，R 为混合检索函数，ScoreAgg 为加权分数累积函数。经测试，在N=10的情况下能够获得较好的稳定性提升，相比更大的采样次数具有更高的性价比。该策略的核心动机在于利用不同融合权重下检索结果的互补性：当某一候选技术在多种权重组合下都获得较高分数时，其累积得分会显著高于仅在个别权重下出现的噪声候选，从而提升最终检索结果的稳定性。

3.5 重排序与生成模块

混合检索返回的 Top-K 候选集合 C={$c_1,\cdots,c_k$} 可能包含噪声。重排序模块 M(C) 利用 LLM 的推理能力对候选文档重新打分与排序。本文采用 pointwise 重排序策略，通过思维链（chain-of-thought，CoT）提示引导模型分析规则与候选技术的相关性，并输出 0 - 100 之间的置信度分数。得分最高的 Top-K′ 个候选构成最终上下文 C′[40]。

生成模块 G(x, C′) 接收原始规则 x 和精排后的上下文 C′，输出预测标签集合 $\hat{Y}$。生成过程定义为：

$$G(x,C')=PostProcess(LLM_{gen}(Prompt(x,C')))$$

其中，LLM_{gen} 为生成式大语言模型（Qwen3 系列），Prompt 将规则与检索到的知识组织为结构化提示，PostProcess 执行标签标准化与去重。提示设计遵循三项原则：其一，任务明确性，明确要求输出 ATT&CK 标签并给出格式示例；其二，知识增强性，将检索到的技术描述作为上下文嵌入提示，使模型基于外部证据而非仅依赖参数记忆进行判断；其三，结构化输出，要求模型以规定的标记格式输出预测结果，便于后处理解析。

3.6 后处理与标签标准化

由于大语言模型的原始输出格式多样且不规范，需要添加后处理模块将非结构化的输出转化为结构化标签。给定LLM原始输出，后处理函数定义为：

$$Y_{pred}=Dedup(Normalize(Extract(Y_{raw})))$$

先对输出逐行分割并分词，筛选以attack.开头的标记，然后为了避免LLM产生的标签粘连，用正则表达式在标识符前面加上分隔符。标准化步骤把不同的格式统一起来，技术标签提取数字ID重组成“attack.txxxx”或者“attack.txxx.yyy”格式，战术标签转换成“attack.{name}”格式（小写、空格替换成连字符）。去重步骤Dedup则通过集合操作移除重复标签。

4 实验与分析

本章用实验来验证RAG-S2A方法的有效性。实验设计以三个主要问题为中心，即检索增强是否能提高映射的准确性、领域适应性对检索性能的影响程度、以及不

同规模模型对检索增强的响应差异。

4.1 数据集与实验设置

本文从SigmaHQ官方GitHub仓库中获取了3 765条Sigma规则，包含Windows、Linux、macOS等各个平台的检测场景。数据集包含MITRE ATT&CK框架的268个不同的ATT&CK标签（战术和技术/子技术），平均每条规则标注2.4个标签。数据集按69.2%、15.6%、15.1%的比例分为训练集（2 607条）、验证集（588条）和测试集（570条），用分层抽样的方法保证所有标签在训练集中至少出现一次[41]。

由于该任务属于多标签分类，给定真实标签集合Y与预测标签集合$\hat{Y}$，采用精确率（Precision）、召回率（Recall）和F1分数作为评价指标：

$$P=\frac{|Y\cap\hat{Y}|}{|\hat{Y}|}R=\frac{|Y\cap\hat{Y}|}{|Y|}F_1=\frac{2PR}{P+R}$$

此外，为评估检索阶段的质量，定义检索准确率Recall@K为：在Top-K检索候选集合$C^{(K)}$中命中的真实标签占比，即$\text{Recall@K}=\frac{|Y\cap C^{(K)}|}{|Y|}$。

检索模型采用微调后的BGE-M3双编码器，学习率$2\times10-5$，批量大小8，训练25个epoch。生成阶段选用三种Qwen3 API模型：Qwen3-0.6B、Qwen3-4B和Qwen3-Max（API版本：2025-09-23）[42]；同时，在本地4B底座上进一步比较四种RAG感知微调策略：Two-Stage、RAFT、RA-DIT和Self-RAG。为避免只与生成式方法比较，实验还补充了采用相同输入表示的TF-IDF + Logistic Regression和BERT多标签分类基线。检索消融部分比较六种检索方法：无RAG基线、BM25关键词检索、原始BGE稠密检索、微调BGE稠密检索、原始BGE混合检索、微调BGE混合检索。除无RAG基线外，其余五种方法均设置“无重排”和“有重排”两种配置，总计11种实验配置。所有模型对比、消融实验和RAG感知微调比较均在完整测试集（570条规则）上进行评估。

4.2 主要实验结果

表2给出了Qwen3-4B在11种检索配置下的细粒度对比，用于分析性能提升的具体来源。与无RAG基线的20.00%相比，BM25和原始BGE只能将F1提升到33%至35%区间；真正带来显著改进的是检索器的领域适配，微调后的稠密检索将F1提升至54.30%，并把检索准确率从原始BGE的19.45%提高到75.25%。相比之下，混合检索结合微调BGE在该组消融中取得52.33%，略低于单一微调稠密检索，但仍明显优于BM25和未适配的稠密检索，说明该任务上的主要增益首先来自语义检索的领域对齐，而词面匹配信号更多起到补充作用。采用相同输入训练的TF-IDF + Logistic Regression和BERT多标签分类基线F1仅为22.13%和26.69%，进一步说明在268类高度稀疏的多标签场景下，单纯依赖分类范式难以充分利用ATT&CK技术描述中的知识。

图2进一步对比了三种Qwen系列API模型、两类分类基线与四种本地4B RAG感知微调策略在统一设置下的F1表现。结果显示，检索增强对小模型的增益显著大于大模型：Qwen3-0.6B、Qwen3-4B和Qwen3-Max的F1分别从16.80%、20.32%和48.73%提升至32.86%、54.76%和58.97%，对应相对增幅约96%、170%

表2 Qwen3-4B模型完整配置性能对比

%

检索方法	精确率	召回率	F1分数	检索准确率
无RAG(基线)	21.20	20.64	20.00	0.00
BM25	31.49	39.48	33.12	16.20
BM25 + 重排	32.74	39.83	34.01	16.20
稠密(原始BGE)	31.35	40.92	33.66	19.45
稠密(原始BGE) + 重排	33.58	42.56	35.46	19.45
稠密(微调BGE)	52.85	61.52	54.30	75.25
稠密(微调BGE) + 重排	50.52	59.63	52.43	75.25
混合(原始BGE)	33.41	42.05	35.24	20.26
混合(原始BGE) + 重排	33.29	42.14	35.13	20.26
混合(微调BGE)	50.37	60.17	52.33	72.75
混合(微调BGE) + 重排	47.11	57.57	49.45	72.75

和21%。从四种4B微调策略看，Two-Stage、RAFT、RA-DIT和Self-RAG分别达到46.91%、56.85%、60.13%和60.59%，其中后两者均超过Qwen3-Max在相同RAG设置下的58.97%，RAFT也与之接近，说明针对任务设计的RAG感知微调总体上比单纯扩大模型规模更具收益。

4.3 组件有效性分析

本节以Qwen3-4B模型为例做消融实验，分析各个组件的作用。

4.3.1 检索组件

图3展示了五种检索配置在不同K值下的Recall@K表现，其中K表示检索阶段返回的候选技术数量。结果显示，领域微调是稠密检索有效性的前提：原始BGE在K=10时的Recall@K仅为15.10%，与BM25的16.00%基本持平；经过对比学习和困难负样本挖掘后，微调BGE将Dense与Hybrid的Recall@10分别提升至75.30%和74.80%。这说明通用语义检索模型难以直接跨越Sigma规则与ATT&CK描述之间的语义鸿沟，而领域对齐能够显著提升候选标签的覆盖能力。另一方面，Hybrid在较大K值下的召回略低于纯稠密检索，例如K=20时79.80%低于81.70%，表明BM25信号能够补充精确匹配，但在较大检索深度下也可能引入一定的排序稀释效应。

如图4所示，领域对齐微调同时提升了检索准确率与端到端F1。稠密检索的检索准确率由原始BGE的19.45%跃升至微调BGE的75.25%，F1由33.66%提升至54.30%；混合检索的检索准确率也由20.26%提升至72.75%，F1由35.24%提升至52.33%。这说明检索阶段的性能改进能够有效传导至生成阶段，验证了领域适配对端到端性能的系统性收益。

4.3.2 重排序组件

表3为重排序对各个检索方法的详细影响。结果显示，针对检索质量低下的方法（BM25、原始BGE）重排序得到一致的性能提升，但是对于检索质量较高的微调BGE方法来说，重排序却导致性能略有

下降。

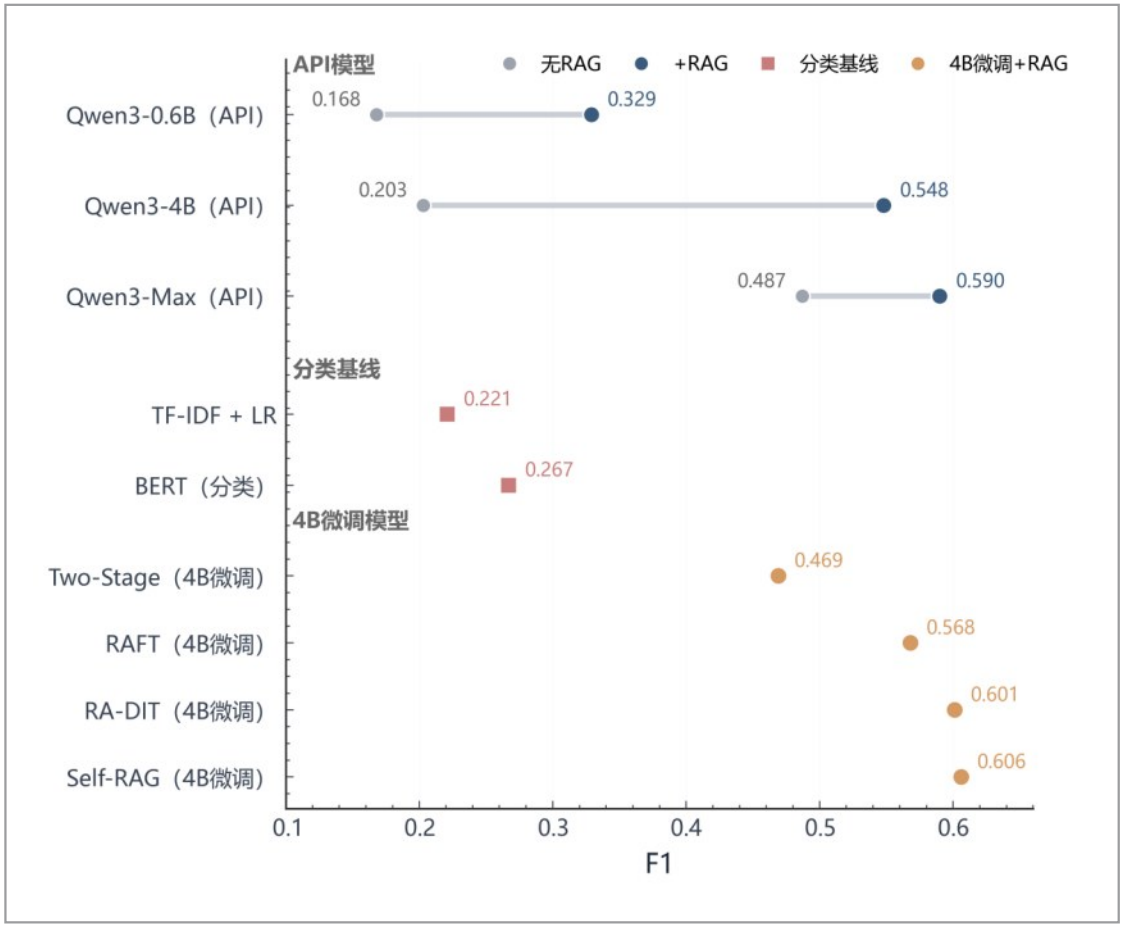


图2　不同规模API模型、分类基线与4B微调模型的F1对比

对于检索质量较低的方法，重排序仍能带来一定帮助：BM25由33.12%提升至34.01%，原始BGE稠密检索由33.66%提升至35.46%，分别提高0.89个百分点和1.80个百分点，说明点式重排序在弱检索条件下能够修正前列结果。

而对于已经完成领域适配的强检索器，重排序的边际收益转为负值：微调BGE稠密检索在检索准确率75.25%时，F1仍由54.30%降至52.43%；混合检索在检索准确率72.75%时也由52.33%降至49.45%。这表明，当初始候选已具备较高质量时，额外重排序未必带来收益，反而可能打乱原有较优排序。

4.3.3　自一致性组件

表4为不同规模模型下自一致性检索策略[43]的性能表现。需要说明的是，该表基于另一组统一的benchmark设置，用于分析自一致性策略的影响，不直接与表2中的检索消融结果进行逐项比较。自一致性检索通过多次独立检索并汇总结果来提高检索的稳定性：对同一个查询做10次重

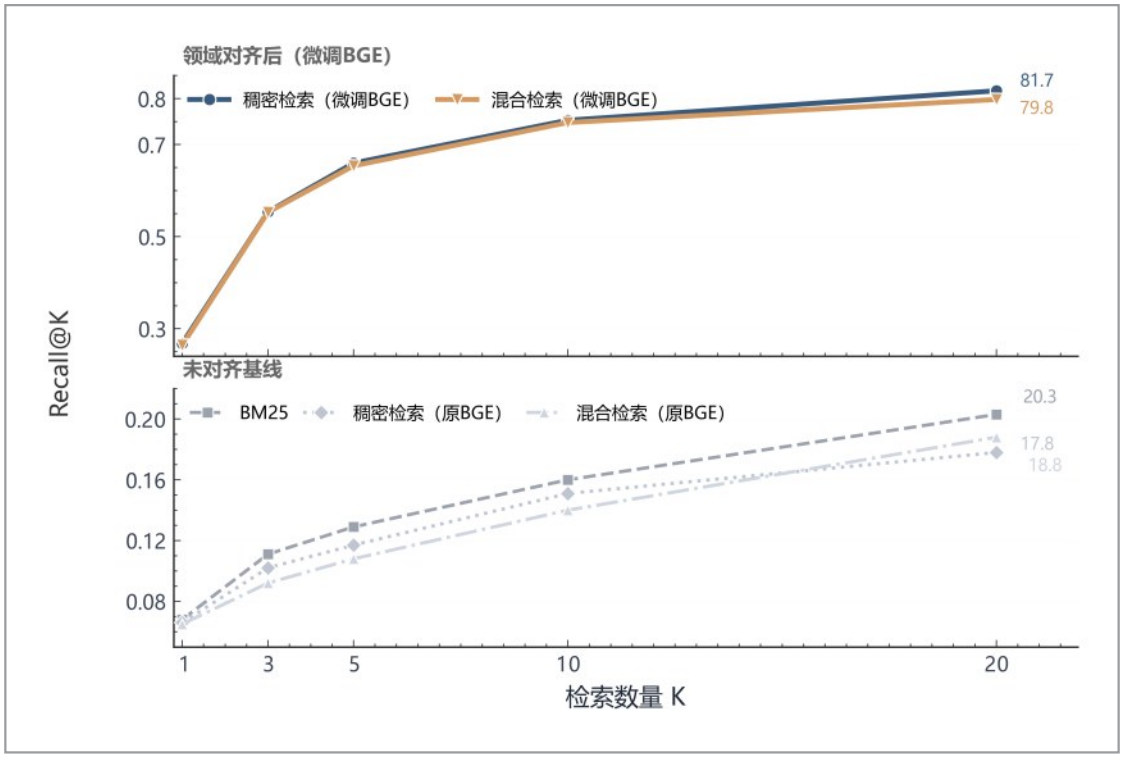


图3 570条Sigma规则测试集上的检索Recall@K对比

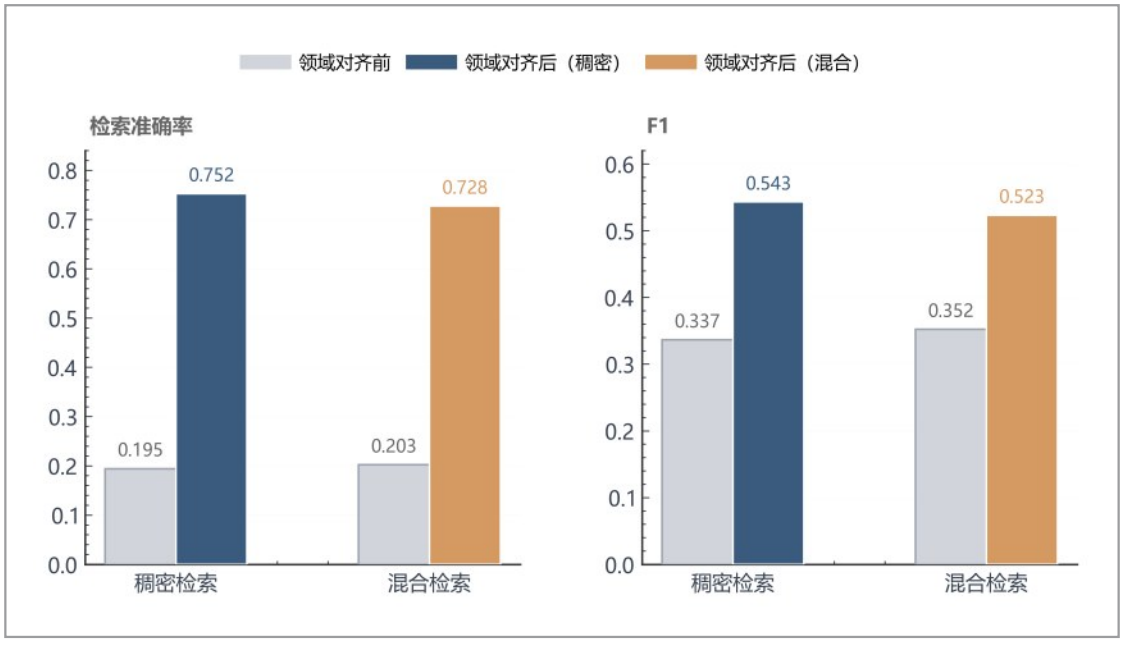


图4 570条Sigma规则测试集上的领域对齐效果对比

复检索，每次采用不同的稀疏–稠密融合权重组合以增加候选多样性，最后通过加

表3 重排序组件效果对比(Qwen3-4B模型,单位:%)

检索方法	无重排（精确率 / 召回率 / F1）	有重排（精确率 / 召回率 / F1）	ΔF1
BM25	31.49 / 39.48 / 33.12	32.74 / 39.83 / 34.01	+0.89
稠密(原始BGE)	31.35 / 40.92 / 33.66	33.58 / 42.56 / 35.46	+1.80
稠密(微调BGE)	52.85 / 61.52 / 54.30	50.52 / 59.63 / 52.43	−1.87
混合(原始BGE)	33.41 / 42.05 / 35.24	33.29 / 42.14 / 35.13	−0.11
混合(微调BGE)	50.37 / 60.17 / 52.33	47.11 / 57.57 / 49.45	−2.88

权分数累积选取得分最高的候选技术作为检索结果。

表4 自一致性组件跨模型效果对比

%

配置	Qwen3-0.6B	Qwen3-4B	Qwen3-Max
无RAG (基线)	20.01	28.23	50.89
混合(微调BGE)	34.07	50.26	53.97
自一致性(微调BGE)	31.13	48.43	55.57
相对变化	−2.94pp	−1.83pp	+1.60pp

实验结果显示出一个显著的“规模依赖”现象：自一致性检索仅对大模型（Qwen3-Max）产生正向增益，而对中小模型（Qwen3-0.6B/4B）则导致性能衰退。这一现象表明，模型能力与集成策略之间存在显著的匹配关系。对于 Max 这样的大模型，其内部知识分布广泛且稳定，多次不同权重下的采样能够捕捉到单一检索遗漏的长尾技术，多样化的候选结果通过加权累积形成了有效的互补，从而充分发挥了集成学习的优势。然而，对于中小模型，由于其推理能力相对有限且输出稳定性较差，多次采样引入的并非有效多样性，而是噪声或不一致的预测；简单的加权累积反而稀释了单次检索中的高质量信号，导致整体性能不升反降。

4.3.4 组件协同与部署建议

综合上述三个组件的消融实验结果，本文提出以下工程实践建议：第一，检索环节是性能提高的前提，应优先保障。从图3的Recall@K分面曲线和图4的检索准确率−F1对比都可以看出，领域适配后的稠密检索把检索准确率从19.45%提升到75.25%，为后续生成环节打下基础。第二，重排序组件要依照需要部署。对于检索质量较低的场景，重排序可以带来一定增益；而面对已具备高质量检索结果的场景，重排序的边际收益很小甚至为负，建议直接跳过该步骤以节省计算资源。第三，自一致性检索更适合大模型场景，该策略增强了大模型在复杂场景下的综合判别能力，但对于中小模型输出不稳定，不宜默认采用。

4.4 案例分析

为了能够对系统的工作过程以及其局限性有更直观的认识，本节选取三个案例进行详细的分析。案例均使用Qwen3-4B模型，用混合检索加上微调BGE的方式。

案例一展示了检索增强机制的有效性及其在低频技术上的局限。在 DNS-over-HTTPS Enabled by Registry 规则中，真实标注包含防御规避战术、T1112（Modify Registry）和 T1140

(Deobfuscate/Decode Files or Information) 三个标签。系统准确识别出防御规避战术与 T1112 技术，检索准确率为 66.67%，推理过程与检索到的技术描述一致，验证了微调后稠密检索模型的语义匹配能力。然而 T1140 技术被遗漏，原因在于该技术在训练集中样本稀少，检索模型未能充分学习其向量表示，导致生成阶段缺少关键上下文。这表明 RAG 系统的性能边界不仅取决于模型的推理能力，也受限于检索模型对低频技术的学习质量。

案例二揭示了标注粒度差异导致的评估偏差。Sigma 规则 Qakbot Regsvr32 Calc Pattern 的真实标注仅包含防御规避和执行两个战术标签，而系统额外预测了 T1218 (System Binary Proxy Execution) 及其子技术 T1218.010 (Regsvr32)，按严格匹配标准被判定为假阳性。然而，模型的推理准确捕捉了规则中 regsvr32 代理执行的语义特征，检索准确率达到 100%，预测结果实际上符合 ATT&CK 框架的层级结构。该案例表明，当评估采用严格匹配标准时，模型在技术层级上的合理预测会因与人工标注粒度不一致而被低估。

案例三展示了技术定义边界模糊所引发的预测错误。以一条包含注册表操作的 Sigma 规则为例，检索模块返回了 T1564 (Hide Artifacts) 和 T1112 (Modify Registry) 两个语义相近的技术描述。由于修改注册表配置本身可以作为隐藏 artifacts 的手段，两者定义存在功能重叠，系统最终将二者混淆并选择了错误的标签。该案例反映出 MITRE 框架中部分技术定义边界不够清晰，即使是人类专家也可能对此类边界案例给出不同判断，这在客观上为模型性能设定了理论上限。

5 结束语

本文针对 Sigma 规则到 MITRE ATT&CK 的自动映射任务，提出了基于检索增强生成的 RAG-S2A 框架，围绕混合检索、领域适配与受控生成三个环节开展了系统实验。实验结果表明，检索增强对中小规模模型的增益远高于大模型，经领域适配的检索模型是性能提升的关键环节，而重排序与自一致性检索等可选组件需根据模型能力与资源条件进行取舍。上述发现为资源受限场景下的模型选型与系统配置提供了实证参考。

但本文方法仍存在以下不足：(1) 标注数据的覆盖度有限，长尾技术因训练样本不足而难以被检索模型有效表征，且不同标注者对标注粒度的理解差异会给模型训练引入噪声；(2) ATT&CK 分类体系中部分技术在功能描述上存在交叉，这种固有的分类边界模糊性构成了自动映射任务的重要难点，并在一定程度上限制了性能上限；(3) 当前检索准确率仍有较大提升空间，检索阶段的召回上限直接制约了生成阶段的性能上限。因此，当前最优 F1 虽然达到 60.59%，但与理想性能之间仍存在差距；不过相较于 BERT 多标签分类基线的 26.69%，RAG-S2A 已显著提升了知识密集型映射任务的实用性，在实际 SOC 场景中可以作为候选标签与检索证据的辅助工具，帮助分析师降低从零标注成本。未来工作将从以下方面展开：第一，引入 ATT&CK 技术关系图谱中的层级结构信息以改善检索召回；第二，探索对生成模型进行领域微调以进一步提升映射精度；第三，将方法推广至其他类型的检测规则与安全知识映射任务。

参考文献：

[1] STROM BLAKE, APPLEBAUM ANDY,

MILLER DOUGLAS, et al. MITRE ATT&CK: Design and Philosophy[R]. McLean, VA: The MITRE Corporation, 2020.

[2] SIGMAHQ. Sigma Rules Specification (v2.1.0) [EB/OL]. 2025. https://sigmahq.io/sigma-specification/specification/sigma-rules-specification.html (accessed 2026-02-04).

[3] CARDINALOPS. State of SIEM Detection Risk: 5th Annual Report (2025 Edition)[R]. Tel Aviv: CardinalOps, 2025.

[4] DEVLIN JACOB, CHANG MING-WEI, LEE KENTON, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019), June 2-7, 2019, Minneapolis, USA. Association for Computational Linguistics, 2019: 4171-4186.

[5] MITRE. Updates - April 2025 (ATT&CK v17) [EB/OL]. https://attack.mitre.org/resources/updates/updates-april-2025/ (accessed 2026-02-04).

[6] LEWIS PATRICK, PEREZ ETHAN, PIKTUS ALEKSANDRA, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), December 6-12, 2020, Online. Red Hook, NY: Curran Associates, 2020: 9459-9474.

[7] ROBERTSON STEPHEN, ZARAGOZA HUGO. The Probabilistic Relevance Framework: BM25 and Beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[8] KARPUKHIN VLADIMIR, OGUZ BARLAS, MIN SEWON, et al. Dense Passage Retrieval for Open-Domain Question Answering[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), November 16-20, 2020, Online. Association for Computational Linguistics, 2020: 6769-6781.

[9] KHATTAB OMAR, ZAHARIA MATEI. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2020), July 25-30, 2020, Online. New York, NY: ACM, 2020: 39-48.

[10] GAO LUYU, MA XUEGUANG, LIN JIMMY, et al. Precise Zero-Shot Dense Retrieval without Relevance Labels[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), July 9-14, 2023, Toronto, Canada. Association for Computational Linguistics, 2023: 1762-1777.

[11] VAN DEN OORD AARON, LI YAZHE, VINYALS ORIOL. Representation Learning with Contrastive Predictive Coding[EB/OL]. arXiv: 1807.03748, 2018. https://arxiv.org/abs/1807.03748 (accessed 2026-02-04).

[12] GAO TIANYU, YAO XINGCHENG, CHEN DANQI. SimCSE: Simple Contrastive Learning of Sentence Embeddings [C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021), November 7-11, 2021, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, 2021: 6894-6910.

[13] CENTER FOR THREAT-INFORMED DEFENSE MITRE). TRAM: Threat Re-

port ATT&CK Mapper[EB/OL]. https://github. com/center-for-threat-informed-defense/tram (accessed 2026-02-04).

[14] SIGMAHQ. sigma_attack_nav_coverage. json (Sigma rule ATT&CK coverage export) [EB/OL]. https://github. com/SigmaHQ/sigma/blob/master/other/sigma_attack_nav_coverage. json (accessed 2026-02-04).

[15] 潘亚峰, 周天阳, 朱俊虎, 等. 基于ATT&CK的APT攻击语义规则构建[J]. 信息安全学报, 2021, 6(3): 77-90.
PAN YAFENG, ZHOU TIANYANG, ZHU JUNHU, et al. Construction of APT attack semantic rules based on ATT&CK [J]. Journal of Cyber Security, 2021, 6 (3): 77-90.

[16] HUSARI GHAITH, AL-SHAER EHAB, AHMED MOHIUDDIN, et al. TTPDrill: Automatic and Accurate Extraction of Threat Actions from Unstructured Text of CTI Sources[C]//Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017), December 4-8, 2017, Orlando, FL, USA. New York, NY: ACM, 2017: 103-115.

[17] ALVES PAULO M M R, FILHO GERALDO P R, GON ALVES VIN CIUS P. Leveraging BERT's Power to Classify TTP from Unstructured Text[C]//Workshop on Communication Networks and Power Systems (WCNPS 2022), November 17-18, 2022, Fortaleza, Brazil. IEEE, 2022: 1-7.

[18] LIU CHENJING, WANG JUNFENG, CHEN XIANGRU. Threat intelligence ATT&CK extraction based on the attention transformer hierarchical recurrent neural network[J]. Applied Soft Computing, 2022, 122: 108826.

[19] ALAM MD TANVIRUL, BHUSAL DIPKAMAL, PARK YOUNGJA, et al. Looking Beyond IoCs: Automatically Extracting Attack Patterns from External CTI[C]//International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2023), October 16-18, 2023, Hong Kong, China. ACM, 2023: 92-108.

[20] MĂRMUREANU MARIUS, OPRIŞA CIPRIAN. MITRE Tactics Inference from Splunk Queries[C]//IEEE 19th International Conference on Intelligent Computer Communication and Processing (ICCP 2023), October 26-28, 2023, Cluj-Napoca, Romania. IEEE, 2023: 277-283.

[21] RANI NANDA, SAHA BIKASH, MAURYA VIKAS, et al. TTPHunter: Automated Extraction of Actionable Intelligence as TTPs from Narrative Threat Reports[C]//Australasian Computer Science Week (ACSW 2023), January 30-February 3, 2023, Melbourne, Australia. New York, NY: ACM, 2023: 126-134.

[22] RANI NANDA, SAHA BIKASH, MAURYA VIKAS, et al. TTPXHunter: Actionable Threat Intelligence Extraction as TTPs from Finished Cyber Threat Reports[J]. Digital Threats: Research and Practice, 2024, 5(4): 1-19.

[23] DANIEL NIR, KAISER FLORIAN KLAUS, GILADI SHAY, et al. Labeling Network Intrusion Detection System (NIDS) Rules with MITRE ATT&CK Techniques: Machine Learning vs. Large Language Models[J]. Big Data and Cognitive Computing, 2025, 9(2): 23.

[24] FAYYAZI REZA, TAGHDIMI ROZHINA, YANG SHANCHIEH JAY. Advancing TTP Analysis: Harnessing the Power of Large Language Models with Retrieval Augmented Generation[C]//Annual Computer Security Applications Confer-

ence Workshops (ACSAC Workshops 2024), December 9–10, 2024, Honolulu, HI, USA. IEEE, 2024: 255–261.

[25] WUDALI PRASANNA N, KRAVCHIK MOSHE, MALUL EHUD, et al. Rule-ATT&CK Mapper (RAM): Mapping SIEM Rules to TTPs Using LLMs[EB/OL]. arXiv: 2502.02337, 2025. https://arxiv.org/abs/2502.02337 (accessed 2026-02-04).

[26] ZHANG YONGHENG, DU TINGWEN, MA YUNSHAN, et al. AttacKG+: Boosting attack graph construction with Large Language Models[J]. Computers & Security, 2024, 150: 104220.

[27] LEKSSAYS AHMED, SHUKLA UTSAV, SENCAR HUSREV TAHA, et al. TechniqueRAG: Retrieval Augmented Generation for Adversarial Technique Annotation in Cyber Threat Intelligence Text[C]//Findings of the Association for Computational Linguistics: ACL 2025, July 27–August 1, 2025, Vienna, Austria. Association for Computational Linguistics, 2025: 20913–20926.

[28] 刘天扬, 寇思佳, 金旭, 等. 基于图检索增强生成和少样本学习的美术作品鉴赏[J]. 大数据, 2025, 11(05): 101–116.
LIU TIANYANG, KOU SIJIA, JIN XU, et al. Art appreciation based on graph retrieval-augmented generation and few-shot learning[J]. Big Data Research, 2025, 11(05): 101–116.

[29] ASAI AKARI, WU ZEQIU, WANG YIZHONG, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[C]//International Conference on Learning Representations, Vienna: OpenReview.net, 2024.

[30] LIN XI VICTORIA, CHEN XILUN, CHEN MINGDA, et al. RA-DIT: retrieval-augmented dual instruction tuning[C]//International Conference on Learning Representations, Vienna: OpenReview.net, 2024.

[31] ZHANG TIANJUN, PATIL SHISHIR G, JAIN NAMAN, et al. RAFT: adapting language model to domain specific RAG [EB/OL]. arXiv: 2403.10131, 2024. https://arxiv. org/abs/2403.10131 (accessed 2026-02-04).

[32] YAN SHI-QI, GU JIA-CHEN, ZHU YUN, et al. Corrective retrieval augmented generation[EB/OL]. arXiv: 2401.15884, 2024. https://arxiv. org/abs/2401.15884 (accessed 2026-02-04).

[33] LIANG XUN, NIU SIMIN, LI ZHIYU, et al. SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), July 27–August 1, 2025, Vienna, Austria. Association for Computational Linguistics, 2025: 4609–4631.

[34] WANG YU, ZHAO SHIWAN, WANG ZHIHU, et al. RAG+ : Enhancing Retrieval-Augmented Generation with Application-Aware Reasoning[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, November 4–9, 2025, Suzhou, China. Association for Computational Linguistics, 2025: 32013–32037.

[35] SOC PRIME. AI-Powered ATT&CK Tag Prediction for Sigma Rules by Uncoder AI[EB/OL]. https://socprime. com/blog/uncoder-ai-automates-mitre-attck-tagging-in-sigma-rules/ (accessed 2026-02-04).

[36] CENTER FOR THREAT-INFORMED DEFENSE MITRE). SigmaGen: AI-Powered Sigma Rules Generation with MITRE ATT&CK[EB/OL]. https://ctid.mitre. org/events/apac-2025/ (accessed

2026-02-04).

[37] BRUCH SEBASTIAN, GAI SIYU, INGBER AMIR. An analysis of fusion functions for hybrid retrieval[J]. ACM Transactions on Information Systems, 2024, 42(1): 1-35.

[38] CHEN JIANLV, XIAO SHITAO, ZHANG PEITIAN, et al. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation[C]//Findings of the Association for Computational Linguistics: ACL 2024, August 11-16, 2024, Bangkok, Thailand. Association for Computational Linguistics, 2024: 2318-2335.

[39] MOREIRA GABRIEL DE SOUZA P, OSMULSKI RADEK, XU MENGYAO, et al. NV-Retriever: improving text embedding models with effective hard-negative mining[EB/OL]. arXiv: 2407.15831, 2024. https://arxiv.org/abs/2407.15831 (accessed 2026-02-04).

[40] WEI JASON, WANG XUEZHI, SCHUURMANS DALE, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models[C]//Advances in Neural Information Processing Systems (NeurIPS 2022), November 28-December 9, 2022, New Orleans, LA, USA. Curran Associates, 2022.

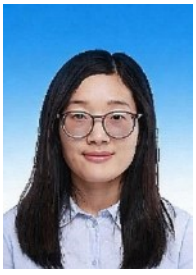
[41] SIGMAHQ. Main Sigma Rule Repository [EB/OL]. https://github.com/SigmaHQ/sigma (accessed 2026-02-04).

[42] QWEN TEAM. Qwen3 Technical Report [EB/OL]. arXiv: 2505.09388, 2025. https://arxiv.org/abs/2505.09388 (accessed 2026-02-04).

[43] WANG XUEZHI, WEI JASON, SCHUURMANS DALE, et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models[C]//International Conference on Learning Representations (ICLR 2023), May 1-5, 2023, Kigali, Rwanda. OpenReview.net, 2023.

作者简介

陈烨楷（2000-），男，华东师范大学数据科学与工程学院硕士生，主要研究方向为网络安全、大数据。



田玉丹（1990-），女，华东师范大学信息化治理办公室中级工程师，主要研究方向为网络安全。

赵明昊（1995– ），男，华东师范大学数据科学与工程学院讲师、晨晖学者，清华大学博士，主要研究方向为计算机系统、数据库系统、数据科学综合应用。

钱卫宁（1976– ），男，华东师范大学数据科学与工程学院教授、博士生导师，院长，复旦大学博士。主要研究方向为可扩展事务处理、大数据管理系统基准评测、海量数据分析处理及其应用。

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: