DOI：10.20189/j.cnki.CN/61-1527/E.202601012

忠实性增强的国防政策检索增强生成方法

罗　云，徐嘉骏，彭明阳*，朱竞丹
（国防科技大学 外国语学院，南京 210039）

摘　要：为解决标准检索增强生成（retrieval-augmented generation，RAG）模型在处理国防白皮书等高权威性、结构化文本时存在的语义偏差、上下文信息丢失及知识安全风险等问题，提出一种融合结构感知与交叉验证的检索增强生成方法（structure-aware and verification-enhanced RAG，SV-RAG）。首先，在知识库构建阶段，采用“结构化切块”方法，在切分文本的同时保留并利用文档的原始层次结构元数据。其次，在知识检索阶段，设计了双轨混合检索机制，针对权威知识库，提出一种查询感知的自适应加权重排序算法，融合文本相似度与元数据匹配度以精确捕捉上下文内涵；针对外部检索信息，则通过“双流交叉核查”流程，以权威知识为基准验证并整合外部信息。最后，在问答生成阶段，设计了“原文引用与精准应答”生成范式，通过提示工程强制模型先引用核心原文再进行阐述，以保障答案的忠实度。为验证该方法的有效性，构建了包含494对高质量问答对的中国国防安全知识问答数据集，在该数据集上的综合实验结果表明：所提出的SV-RAG架构在忠实度和语境准确性等核心指标上表现显著，相较于传统RAG基线，性能分别提升28.0%和28.2%。

关键词：大语言模型；事实忠实度；知识检索；权威文本
中图分类号：TP391　　**文献标志码**：A　　**开放科学标识码**（OSID）：
文章编号：2097-2741（2026）01-0132-11

听语音　聊科研　与作者互动

国防政策是国家安全战略的基石，其官方文件具有内容高度权威性、结构严谨性和复杂性。如何从这些高密度文本中快速、准确地获取信息，并确保对政策内涵的理解不产生偏差，是国防领域信息服务面临的关键挑战。在此背景下，构建面向国防政策的智能问答系统，不仅能提升信息检索效率，还可以在新信息环境下保障政策解读的权威性，并以此辅助战略决策、服务国防教育和引导公众认知。

信息检索与问答系统（question answering，QA）的发展为解决上述挑战提供了可行路径。早期系统多依赖关键词匹配与人工规则，难以深入理解自然语言语义[1-2]。随着以GPT为代表的大语言模型（large language models，LLMs）的兴起，问答领域迎来了重大变革[3]。LLMs凭借其强大的语言理解与生成能力，能够提供流畅的回答，但其固有的“事实幻觉”[4]与知识更新滞后等缺陷，使其在直接应用于高权威性场景时存在巨大风险。为应对这些瓶颈，检索增强生成（retrieval-augmented generation，RAG）应运而生，其通过结合外部知识库提升答案的准确性，迅速成为主流技术范式[5]。

尽管标准RAG框架在通用领域取得了巨大成功[6-7]，但将其直接应用于国防政策文件这类高度权威、结构化且敏感的文本时，仍存在明显局限性。近年来，有研究尝试将检索生成方法应用于其他领域，如法律文书分析[8-9]、医疗诊断问答[10]等，尽管这些领域与国防政策领域相似，同样要求极高的准确性，但面对国防安全领域特有的权威性、结构性和安全性三重挑战的综合性研究仍

收稿日期：2025-08-02　　**修回日期**：2025-10-10　　**录用日期**：2025-10-22
第一作者：罗云（2002—），男，硕士研究生，主要从事语言信息处理研究。E-mail：1315494900@qq.com
***通信作者**：彭明阳（2002—），男，硕士研究生，主要从事情报分析研究。E-mail：2386896109@qq.com

引用格式：罗云，徐嘉骏，彭明阳，等.忠实性增强的国防政策检索增强生成方法[J].火箭军工程大学学报，2026，40（1）：132-142.
LUO Yun，XU Jiajun，PENG Mingyang，et al. Faithfulness-enhanced retrieval-augmented generation method for national defense policies[J]. Journal of Rocket Force University of Engineering，2026，40（1）：132-142.

显不足，现有方法的局限性主要体现如下。

1）生成内容忠实度不足。标准RAG的生成模块本质上是释义而非忠实复述，其固有的“创造性”在面对措辞严谨的国防政策时，极易引入语义偏差，导致答案与官方立场不一致，影响内容的权威性[11]。

2）文档结构信息利用缺失。政策文件具有清晰的章节结构，同一术语在不同章节的语境下具有不同内涵。传统RAG的处理方式忽略了这种层次结构，导致上下文信息丢失，无法精确理解特定语境下的政策意涵[12]。

3）外部信息验证机制缺乏。为增强时效性与全面性，RAG常需引入外部信息。但在高敏感的国防领域，通用方法缺乏以权威知识为基准的验证流程，无法有效过滤外部信息中的潜在错误、虚假或过时内容，这为国防建设带来了严重的安全隐患[13]。

为应对上述三大挑战，着重从提升生成内容忠实度、增强结构化信息利用和建立外部信息验证机制3个方面进行系统性优化，提出了一种融合结构感知与交叉验证的检索增强生成方法（structure-aware and verification-enhanced RAG，SV-RAG），旨在构建一个可靠、精确的国防政策智能问答系统。与现有研究成果相比，本文并非只是对RAG的某一单点进行优化，而是从知识库构建、知识检索到问答生成的全流程进行针对性设计，形成了一套专门面向高权威性、结构化文本的综合解决方案。

1 理论基础

本研究工作与RAG、高级检索与重排序技术以及生成内容的事实忠实度保障等领域密切相关。

1.1 RAG

RAG的核心思想是将参数化的语言模型与非参数化的外部知识库相结合[5]。标准RAG流程包括“检索”与“生成”两步。首先，利用检索器从知识库中召回与用户问题相关的文档片段；然后，将其与原始问题一同输入给生成器以产出最终答案。这一范式有效改善了LLMs在处理知识密集型任务时的幻觉问题，并允许系统利用最新或领域特定的知识[14]。

随着RAG技术的发展，研究者们从不同角度对其进行了优化。例如，在检索阶段，从早期的TF-IDF、BM25等稀疏向量检索[15]，发展到当前主流的基于稠密向量的语义检索（如DPR）[2]。在整体架构上，也出现了如Self-RAG[16]、Corrective-RAG[17]等更为复杂的框架，通过引入反思、修正等机制来提升RAG流程的鲁棒性。然而，这些通用的RAG改进方案大多将知识源视为无差别的“文本块”集合，忽略了文档固有的结构化信息。对于法律条文、技术手册、政策文件等高度结构化的文本，对文档结构信息的忽视会严重阻碍系统对上下文的深度理解[18]。对此，在SV-RAG中设计“结构感知”模块，着力解决这一核心问题。此外，在军事应用领域，已有研究者针对特定场景开展了深入的信息处理工作，例如，采用神经网络模型从航母情报文本中进行关系抽取[19]，同样凸显了在特定领域内进行精细化、结构化信息处理的重要性。

1.2 高级检索与重排序技术

为了提升RAG中检索环节的“查准率”，研究者们探索了多种高级检索策略。混合检索是其中一种有效方法，它结合了关键词检索（如BM25）在精确匹配上的优势和语义检索在理解相关性上的优势[20]。另一种广泛应用的技术是重排序，该技术在初步召回一批候选文档后，使用一个更强大但计算成本也更高的模型，通常为交叉编码器Cross-Encoder[21]，对候选文档与问题的相关性进行二次打分和排序。尽管这些技术提升了检索的整体相关性，但其评分的依据仍然局限于查询与文本块内容的语义相似度，未能充分利用文本块之外的元数据信息，如其所属的章节标题。在国防政策文本中，章节标题是理解其下属内容核心主旨的关键。对此，提出“查询感知的自适应加权重排序”算法，对现有重排序思想进行扩展，将文本相关性与上下文匹配度进行加权融合，使检索过程能够感知并利用文档的宏观结构。

1.3 生成内容的事实忠实度

确保生成内容与源知识的一致是RAG乃至所有LLMs应用的核心挑战之一[22-23]。为缓解“事实幻觉”，研究者们进行了诸多探索。一方面，通过优化检索内容，提供更高质量、更相关的上下文[24]；另一方面，在生成阶段对模型进行约束。主要包括：知识溯源与引用，即让模型在生成内容的同时，明确指出其信息来源的原文片段[25]；事实核查与修正，即利用独立模型或LLMs自身能力对生成内容进行后处理校验[26]；通过精巧的提示工程引导模型生成更忠实的回答[27]。

现有方法在提升事实性上取得了显著进展，

但对于国防政策这类措辞极其严谨的领域，即使是经过释义的回答，也可能因未能传达原文的精确立场和语气而产生误导。同时，构建可解释的知识系统也是保障信息可靠性的另一重要途径。例如，有研究综述了如何利用知识图谱推理的解释方法，通过提供清晰的推理路径来增强系统的透明度和可信度[28]。为综合提高生成内容的事实性与解释忠实性，提出“原文引用与精准应答”生成范式，以严格保障忠实度。其通过在输出结构上强制模型先“引用”后“回答”，将答案的权威性“锚定”在不可篡改的原文之上，从根本上约束了模型的自由发挥，是保障国防政策问答系统权威性的关键措施。

2 SV-RAG模型架构

基于上述分析，对传统RAG范式进行优化，提出SV-RAG以满足国防安全领域的特殊需求。其构成包含知识存储、知识检索与问答生成3个核心模块，整体流程如图1所示。

针对国防政策文件所特有的权威性与结构化特征，在各模块中融入了针对性改进。一是在知识存储阶段，引入结构化元数据以保留文档的原始层次信息；二是在知识检索阶段，设计了一种面向权威知识库的查询感知的自适应加权重排序（query-aware adaptive weighting reranking，QA-AWR）算法，并辅以一种针对网络信息的跨源验证与摘要机制；三是在问答生成阶段，采用了一种遵循“原文引用与精准应答”原则的生成策略，以确保答案的高忠实度与可读性。

2.1 融合结构化元数据的知识库构建

为解决传统知识库构建中因忽略文档结构而导致的上下文信息丢失问题，针对国防白皮书章节分明、逻辑严谨的特点，构建了一个能精确反映其内在层次结构的机器可读知识库，以确保对政策的理解能够保留其预设语境。首先，进行数据预处理，通过解析白皮书原文，将文本、图表等内容统一为文本格式。然后，采用有别于传统的固定长度切块方法的“结构化切块”，在文本分割时同步提取并保留每个文本块的结构化元数据。最后，利用BCEmbedding模型（https://github.com/netease-youdao/BCEmbedding/blob/master/README_zh.md）将每个文本块映射为向量表示，并与其元数据一同存入向量数据库。该存储结果不仅保留了文本的语义信息，更重要的是嵌入了其在原始文档中的层次结构与宏观语境信息，为后续的高精度检索奠定了坚实基础[14]。

2.2 双轨混合检索与验证机制

为同时应对国防领域术语的上下文强依赖性与外部信息应用的严格安全性两大挑战，知识检索模块设计了一套双轨混合检索与验证机制，旨在通过内部的元数据语义重排序解决术语歧义问

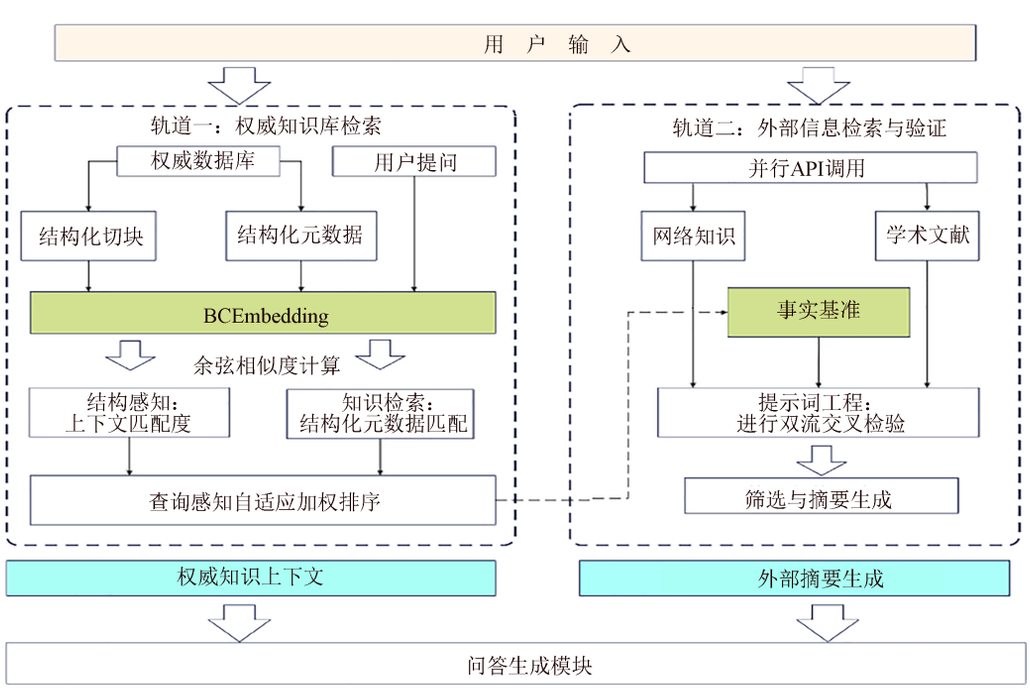

图 1 SV-RAG模型架构

Fig. 1 Architecture diagram of SV-RAG model

题，并通过外部的双流交叉核查措施确保新增信息的权威性与安全性，从而在保证精确性的前提下实现信息的有效增广。

2.2.1 QAAWR

为缓解国防术语的歧义性并实现对用户意图更精确的理解，传统方法通常采用固定权重线性加权的方式来融合文本相关性与上下文相关性，但这需要人工设定超参数，且无法适应不同类型查询的需求，而QAAWR算法能够根据查询的特性对权重进行动态调整，无需任何预设超参数，流程如下。

针对用户问题，通过BCEmbedding模型进行标准的向量相似度计算。BCEmbedding模型将用户的查询转换成向量$\boldsymbol{V}_{\mathrm{q}}$，知识库中每一个文本块的向量$\boldsymbol{C}$是预先计算好并存储的，同样使用BCEmbedding模型，将这个结构化元数据字符串也转换成一个向量$\boldsymbol{S}$。

从知识库中召回Top-K个候选文本块。此阶段，获得了每个候选块的文本相关性T和上下文匹配度E，分别可表示为

$$T_i=\cos\left(\boldsymbol{V}_{\mathrm{q}},\boldsymbol{C}_i\right) \tag{1}$$

$$E_i=\cos\left(\boldsymbol{V}_{\mathrm{q}},\boldsymbol{S}_i\right) \tag{2}$$

式中：$\boldsymbol{V}_{\mathrm{q}}$为用户查询的向量表示；$\boldsymbol{C}_i$为某一文本块的向量表示；$T_i$表示$\boldsymbol{V}_{\mathrm{q}}$与$\boldsymbol{C}_i$之间的余弦相似度；$\boldsymbol{S}_i$为结构化元数据向量；$E_i$表示$\boldsymbol{V}_{\mathrm{q}}$与$\boldsymbol{S}_i$之间的余弦相似度。

算法的核心在于利用初步召回结果的特性来动态生成权重。该算法的基本假设是，初步召回的最高文本相关性得分，可以作为查询具体性的一个有效指标。首先，在K个候选块中找到最高的文本相关性得分M，可表示为

$$M=\max\left(T_1,T_2,T_3,T_4,T_5\right) \tag{3}$$

然后，将M直接作为动态权重，用于计算每个候选块的最终得分F_i为

$$F_i=M\cdot T_i+(1-M)\cdot E_i \tag{4}$$

如果用户的查询非常具体，能够找到一个文本匹配度极高的候选块，即M趋近于1，则权重将偏向文本相关性，确保这个最佳匹配能够获得最高排序。反之，如果查询较为宽泛，所有候选块的文本匹配度都有限，即M较低，那么算法将自动增加上下文匹配度的权重，依靠章节结构信息来辅助判断用户的真实意图。最后根据F进行重新排序，选取Top-N个最相关的文本块作为权威知识库的检索结果。

2.2.2 基于双流交叉核查的摘要生成

为安全、可靠地整合源自互联网的实时信息与专业文献的深度分析内容，引入一套精密的外部信息检索与验证机制。该机制利用具备网络与专业文献检索能力的第三方API，并采用“网络搜索”与“文库搜索”2种模式进行双流并行处理。其核心目标在于将多样化的外部信息安全地“锚定”在检索出的权威知识基准之上，从而确保最终信息的可靠性与全面性。该流程具体步骤如下。

Step1 双流并行检索。在执行权威知识库检索的同时，系统向第三方API发起并行请求：其一，通过网络搜索获取与用户问题相关的Top-5条实时网页信息；其二，通过文库搜索获取5篇相关的专业报告或学术论文。为提升初始召回质量，系统在接口调用时启用了“通过网页摘要信息进行召回增强”的设置，以利用高质量摘要优化检索的相关性。

Step2 双流交叉核查。系统进入保障外部信息质量的核心环节，以检索出的权威知识库内容作为“事实基准”，对上述2个信息流进行独立且并行的验证。对于网络搜索结果，系统利用LLMs的事实核查能力[12]，将5条信息片段与事实基准比对，并将其分类为“一致”“新增”“冲突”或“无关”。对于文库搜索结果，系统执行完全相同的验证流程，确保其专业观点与数据同样经过权威基准的审核[14]。

Step3 信息筛选与摘要生成。根据验证结果对2个信息流进行筛选处理：从网络搜索与文库搜索结果中，分别筛选出所有被判定为“一致”和“新增”的信息。最后，系统将这2个经过筛选的信息集合分别加以整合，利用LLMs生成2份独立的、高度浓缩的摘要：一份是“网络实时信息摘要”，总结与权威知识一致的最新网络动态；另一份是“外部文库洞察摘要”，总结可作为补充的深度分析与研究观点。

通过此双流验证与摘要机制，本系统不仅能主动过滤外部信息中的潜在错误，更能结构化地区分和处理不同性质的信息来源（例如，实时新闻与深度文献），为最终的问答生成提供了更为丰富、可靠且多维度的上下文支持。

2.3 遵循“原文引用与精准应答”的生成策略

为有效应对LLMs在释义时固有的“创造性”与国防白皮书措辞“权威、严谨”这一核心矛盾，问答生成模块对传统的开放式生成范式进行了约束，规避可能导致政策误读的语义偏差风险，确

保最终答案在措辞和内涵上均与源文件保持高度一致。遵循“原文引用与精准应答”原则的结构化生成策略利用精密的提示工程[13]，引导大语言模型以一种高度受控的方式构建答案，以确保其输出结果在内涵与措辞上均与官方文件保持高度一致。其核心是一个为本系统专门定制的提示词模板，明确规定了模型的思考路径和输出结构，关键思想如表1所示。

表 1 “原文引用与精准应答”提示词核心思想

Table 1 Core ideas of the prompt words for original text citation and precise response

模块	说明
角色定义	要求模型扮演一个严谨、精确的中国国防政策分析专家，其核心职责是基于白皮书权威内容进行回答
核心任务	明确指令模型必须遵循“措辞的权威性与严谨性”这一根本原则
执行步骤	在检索到的权威信息中，定位最能直接、权威、完整地回答用户问题的“核心引文”；基于该引文以及其他上下文信息进行精准应答
输出格式	强制规定答案必须包含2个严格区分的部分：1. 核心引文：要求一字不差地引用原文；2. 问题回答：要求所有阐述都必须服务于引文，不能曲解或超越引文的核心立场
约束条件	强调引文的“绝对忠实性”，禁止任何形式的修改；同时要求“问题回答”部分必须与引文在逻辑和立场上保持高度一致

在实际生成过程中，系统会将经过元数据语义重排序的权威文本块以及经过跨源验证的“验证后摘要”作为上下文，并应用上述提示词模板引导大语言模型进行最终生成。

该生成策略通过引导模型锚定于一个不变的“核心引文”，为答案奠定了坚实的权威基石。由此，随后的“问题回答”部分并非简单的解释，而是在此基石之上，利用系统提供的丰富、高质量上下文，进行服务于核心引文、必要且精准的阐述。此结构化的生成策略，规避了因模型自由生成而导致的语义偏差风险，从而提升系统在解读国防政策输出时的精确性、权威性与忠实度。

3 数据集构建

为了构建高质量的评测数据集，设计数据构建流程。首先，在数据搜集阶段，将中国国防部白皮书中的图表内容转化为文本格式，并同步从中国国防部、中国政府网等权威平台收集相关的新闻及专家解读作为补充材料。然后，进入数据预处理阶段，对搜集到的全部数据进行细致清洗，具体操作包括剔除重复或相似的问答、实现文本格式标准化以及进行同义词替换，旨在保障数据的高度一致性与品质。在此过程中，文本还严格筛选，剔除了超出国防知识范围的数据，确保每个问题都聚焦于核心领域。最后，在问答对整合阶段，将经过彻底清洗的数据整理成规范的问答对，每一对均包含一个问题及其对应的权威答案。通过结合白皮书原文特征进行问题与答案的精心设置，确保问答对的专业性与评测价值。

通过上述流程，构建了由494个中国国防安全知识问答对组成的数据集，问答对数据集的部分样例如表2所示。

表 2 问答对数据集的部分样例

Table 2 Partial samples of question-answer dataset

序号	问题	答案
1	中国国防费的增长原则和使用方向是什么，如何确保国防费的透明与高效利用	中国国防费坚持适度、合理、可持续的增长原则，与国家经济发展和国防需求相适应。国防费主要用于人员待遇、训练保障、装备研发与采购维护、国防改革、军民融合及国际安全行动。中国定期向联合国提交军费报告，通过国防白皮书公开相关情况，确保国防费被透明高效利用
2	中国推动军民融合深度发展的核心目标、主要领域和实施路径是什么	中国推动军民融合深度发展的核心目标是形成全要素、多领域、高效益的融合发展格局，实现国防建设与经济建设协调平衡兼容发展。主要领域涵盖基础设施、装备科研、关键技术、人才培养、军队保障、国防动员等。实施路径是依托国家统一领导机制，完善政策法规，促进军地资源优化配置
3	中国军队参与国际安全合作的主要形式和核心宗旨是什么，如何为构建人类命运共同体贡献力量	中国军队参与国际安全合作的形式包括联合国维和、亚丁湾护航、国际救援、中外联演联训、军控裁军及军事学术交流。核心宗旨是维护世界和平、促进共同发展，恪守不干涉别国内政、尊重国家主权和联合国宪章原则。中国军队通过上述行动履行大国责任，构建新型安全伙伴关系，为人类命运共同体提供安全保障

4 实验与结果分析

设计对比与消融实验，从多个维度评估SV-RAG在处理权威性、结构性、敏感性文本时的性能，尤其是在答案的忠实度、语境准确性以及由外部信息融合带来的信息完整度提升效果方面，验证其在国防安全领域的有效性。

为回应上述实验目标，本次实验所采用的数据集本身即是上述文本特性的直接体现。

权威性：实验数据核心源自《新时代的中国国防》等官方白皮书，其措辞、立场和数据均代表国家权威，对解读的准确性要求极高，直接对应对答案忠实度的评估需求。

结构性：国防白皮书具备清晰的“章-节”层次结构，同一术语（如“积极防御”）在不同章节的语境下有不同侧重。这使得利用文档结构进行语境准确性评估成为可能，用以检验模型是否能克服传统RAG的“扁平化”处理缺陷。

敏感性：国防政策涉及国家安全，领域高度敏感。这要求系统在利用外部信息时必须具备验证能力，以过滤潜在的错误或误导性内容，与所设计的双流交叉核查机制及后续的忠实度评估直接相关。

4.1 实验设置

4.1.1 实验环境

所有实验均在表3所示的硬件与软件环境中进行。为在信息检索的全面性与计算效率之间取得平衡，并充分考虑所用8b级别模型的上下文窗口限制，实验中的关键检索与召回参数（Top-K，Top-N）均设置为5。

4.1.2 基线对比

为验证SV-RAG架构中各项技术创新的有效性，设置了以下基线用于对比。

1）Naive RAG：一个标准的RAG实现模块。采用传统的固定长度切块方法处理白皮书原文，使用标准的向量相似度检索（Top-5），并采用通用的问答提示词进行生成，不包含任何结构感知、验证或特殊生成策略。

2）SV-RAG消融设置: 为验证本架构各核心模块的贡献，本文设计了一系列消融变体进行对比。包括：SV-RAG w/o Structure移除了结构化切块与结构化元数据重排序，替换为Naive RAG的切块和检索方式；SV-RAG w/o Verification移除外部信息的双流交叉核查模块，直接将外部检索内容与权威知识库内容合并；SV-RAG w/o Quotation移除“原文引用与精准应答”生成策略，采用标准的开放式问答提示词。

表3 参数配置

Table 3 Parameter configuration

类别	项目	具体配置
硬件环境	服务器平台	AutoDL V48G 服务器
	CPU	12核 Intel Xeon(R) Gold 6459C
	系统内存	72 GB
	GPU(核心计算单元)	vGPU(48 GB 显存)
软件环境	操作系统	Linux ubuntu 22.04
	CUDA 版本	11.8
	编程语言	Python 3.10
	主要框架	PyTorch 2.1.2
核心模型	Embedding 模型	bce-embedding-base_v1
	大语言模型	Qwen 3-8b[29], Llama 3.1-8b[30]
	网络检索 API	Metaso Search API (metaso. cn/search-api/playground)
关键参数	权威知识库召回K	5
	重排序后送入文本块N	5
	外部信息检索召回	5

4.1.3 评价指标选取

由于国防政策解读的严肃性，对答案的评估不仅要看其流畅度和相关性，更要关注其事实的准确性与对原文立场的忠实度。因此，采用人工评估与大模型辅助的自动评估相结合的方式。

1）人工评估指标。由3位专家以独立方式进行，对每个系统生成的答案按照忠实度、语境准确性、信息完整度、相关性与可读性4个维度进行打分。评分采用李克特五级量表（1～5分），分数越高代表表现越好，具体的评分细则如表4所示。

2） 自动评估指标。为克服传统自动评估指标在评估事实忠实度与语义准确性方面的局限性，并规避大语言模型绝对评分中的潜在偏见，采用一种基于综合排序的大模型辅助评估方法[16]，旨在提升评估结果的稳定性与区分度。针对测试集中的每个用户问题，将各待评系统生成的答案列表（乱序排列）与标准答案及原始问题整合后，一并输入GPT-4o-mini[31]模型。通过提示工程指导该模型扮演严谨的“事实核查员”，依据忠实度、语境准确性、信息完整度、相关性与可读性4个维度对答案进行综合排序，并为该排序提供详

表 4 人工评估分级规则

Table 4 Rules for manual evaluation grading

分数	忠实度	语境准确性	信息完整度	相关性与可读性
5 分	完全忠实：答案严格遵循知识库原文的立场和事实，无任何信息扭曲、夸大或主观臆测，准确反映官方立场	精准无误：完美且精确地理解术语在白皮书特定章节下的特定内涵，而非宽泛的通用解释，上下文定位精准	完全覆盖：全面覆盖标准答案中的所有核心要点，信息无任何关键性遗漏	高度相关且流畅：回答与问题完全相关，直击要点，逻辑清晰，语言流畅，结构严谨易读
4 分	基本忠实：答案绝大部分内容忠实于原文，但可能存在极轻微的、不影响核心立场的释义偏差	基本准确：基本准确地理解术语的上下文含义，但可能在细节的精确度上略有欠缺，或对上下文的指涉不够具体	基本完整：覆盖绝大部分核心要点，但可能遗漏了个别次要的细节信息	高度相关，偶有冗余：回答主体与问题高度相关，但可能包含少量非核心的冗余信息；整体表达流畅
3 分	存在偏差：答案在核心事实上保持正确，但存在一定程度的释义，导致部分信息与原文立场有轻微出入，或包含非关键性事实错误	理解宽泛：未能完全区分术语在特定上下文中的精确含义，给出了一个较为宽泛或通用的解释，丢失了特定语境下的细节	要点缺失：遗漏了部分核心要点，导致答案不够完整，未能全面回应问题	基本相关，不够聚焦：回答基本与问题相关，但不够聚焦，关键细节不足；语言尚可，但结构可能较为松散
2 分	事实错误：包含明显的事实错误或对原文立场的显著曲解，存在较为明显的主观臆测	理解错误：严重混淆或错误理解术语在不同上下文中的含义，上下文定位出现明显错误	严重缺失：仅覆盖极少数要点，信息严重不完整，无法有效回答问题	关联较弱，表达不清：回答与问题关联较弱，大部分内容离题；或语言表达不通顺，逻辑混乱
1 分	严重失实：回答内容与事实完全不符，存在严重的信息捏造（幻觉），完全违背原文立场	完全错误：完全没有理解问题的上下文，回答与特定背景毫无关联	毫无信息：完全没有回答出任何问题的核心要点	完全无关，无法理解：回答与问题完全无关，或语言表达无逻辑、完全无法理解

尽的论证。为实现量化评估，将排名 R（$R=1, 2, \cdots, Q$，Q 为待评答案总数）映射为数值分数 $L=Q-R+1$。最终，通过计算每个系统在整个测试集上的平均排名分，评估其相对综合性能。

这种基于相对比较的评估范式优势在于强制评估模型进行直接、细致地横向比较，而非给出孤立的绝对分数。不但可以有效降低因模型评分标准浮动带来的不一致风险，而且能够更有效地放大系统间的细微性能差异，从而使最终的量化评估结果与人类专家的偏好展现出更高的一致性。

4.2 实验结果与分析

4.2.1 对比实验

为了系统性地评估SV-RAG架构的有效性及其与不同LLM的协同能力，设计了对比实验。比较所提SV-RAG与标准Naive RAG的性能差异，考察两者分别搭载Qwen 3-8b和Llama 3.1-8b 2款业界主流模型时的具体表现。

1）人工评测。选择3位具备国防领域知识背景的专家，对全部494条问答数据以独立打分方式进行“五分制”评分，具体实验结果如表5所示。

从人工评估结果可以明显看出，所提的SV-RAG架构在提升系统性能方面发挥作用明显。无论基于何种大语言模型，采用SV-RAG架构的系统在各项指标上的得分均显著高于Naive RAG基线。尤其在忠实度和语境准确性这两个核心指标上，SV-RAG展现出显著优势。以Qwen 3-8b模型为例，其提升幅度分别达到了28.0%和28.2%。在所有评估维度上，Qwen 3-8b相较于Llama 3.1-8b表现出微弱但一致的性能优势。一种可能的解释是，Qwen 3-8b在预训练阶段接触了更大规模的中文语料，从而使其在理解与生成中文特定领域文本时更具优势。然而，这种模型间的差距远小于架构优化带来的提升，进一步印证了SV-RAG架构的创新性和核心价值。

表 5 不同模型和架构的人工评估结果

Table 5 Manual evaluation results based on different models and architectures

配置	忠实度	语境准确性	信息完整度	相关性与可读性
Naive RAG+Llama 3.1-8b	3.62	3.58	3.81	4.15
Naive RAG+Qwen 3-8b	3.71	3.65	3.85	4.20
SV-RAG+Llama 3.1-8b	4.68	4.61	4.42	4.55
SV-RAG+Qwen 3-8b	4.75	4.68	4.48	4.59

2）自动评估结果。为进一步验证和量化性能差异，采用了基于综合排序的大模型辅助评估方法[16]。针对测试集中的每个用户问题，将所有待评估系统生成的答案汇总，并连同用户问题和标准答案一并输入给GPT-4o-mini。要求GPT-4o-mini扮演严谨的“事实核查员”，根据答案与标准答案的总体一致性、事实准确性和信息完整性进行综合考量，对列表中的所有答案进行从最佳到最差的排序，并给出理由。将排名结果转化为数值分数（排名第一得4分，第二得3分，以此类推），最终计算每个系统的综合平均排名评分，结果如表6所示。

表6 基于大模型辅助评估的对比实验结果

Table 6 Comparative experimental results based on large-model-assisted evaluation

配置	排名评分
Naive RAG+Llama 3.1-8b	1.85
Naive RAG+Qwen 3-8b	1.96
SV-RAG+Llama 3.1-8b	3.21
SV-RAG+Qwen 3-8b	3.35

大模型辅助评估的结果与人工评估的结论一致，从另一维度验证了SV-RAG架构的显著优越性。在综合排名得分上，SV-RAG架构显著高于Naive RAG，表明其生成结果在评估模型的排序标准下具有更高的综合质量。其中，SV-RAG+Qwen 3-8b的组合表现最佳，获得了3.35的最高平均排名分，相较于Naive RAG + Llama 3.1-8b，性能提升幅度接近81.1%。量化结果证明了SV-RAG架构具备在实际应用中提供更高质量、更可靠的国防政策问答服务的潜力与能力。

4.2.2 消融实验

为精确评估SV-RAG架构中各核心创新模块的独立贡献及其协同效应，基于在对比实验中表现最佳的Qwen 3-8b模型进行了一系列消融实验。通过逐一移除某单一模块观察系统在关键性能指标上的变化。

1）人工评估结果。考虑到实验目的，人工评估结果重点关注忠实度与语境准确性这2个最能反映架构核心优势的指标，结果如表7所示。可以看出，移除结构感知模块 （w/o Structure） 导致语境准确性降低28.4%，凸显了精确捕捉国防白皮书层次结构对于理解特定术语的至关重要性。移除验证增强模块（w/o Verification）后，系统的忠实度降低14.7%，验证了增强模块作为系统的事实核查机制在保障答案可靠性方面的价值。值得注意的是，对忠实度指标影响最为显著的消融操作是移除引用生成模块（w/o Quotation），该操作致使忠实度降低28.0%，充分证明该生成范式是约束LLMs、缓解“事实幻觉”并确保答案权威性的关键机制。

表7 基于Qwen 3-8b的消融实验人工评估结果

Table 7 Manual evaluation results of ablation experiments based on Qwen 3-8b

实验组	忠实度	语境准确性
完整SV-RAG	4.75	4.68
SV-RAG w/o Structure	4.71	3.35
SV-RAG w/o Verification	4.05	4.64
SV-RAG w/o Quotation	3.42	4.62

2） 自动评估结果。同样，利用大语言模型进行了消融实验的自动评估，以综合平均排名分来量化各模块的贡献。结果如表8所示。

表8 基于Qwen 3-8b的消融实验综合平均排名分

Table 8 Comprehensive average ranking scores of ablation experiments based on Qwen 3-8b

实验组	综合平均排名分
完整SV-RAG	3.35
SV-RAG w/o Structure	2.58
SV-RAG w/o Verification	2.93
SV-RAG w/o Quotation	2.55

大模型辅助评估的结果与人工评估的结论表现出高度一致性。相较于完整SV-RAG架构，移除任一核心模块均会导致系统性能的显著下降。具体来看，移除w/o Quotation和w/o Structure模块所带来的性能降幅相对较大，分别为23.9%和23.0%，间接表明答案的忠实度与语境准确性是影响综合质量的核心要素。此外，移除w/o Verification模块同样导致得分明显下降，验证了过滤外部不可靠信息对于生成高质量答案的必要性。

综上，2种评估方法的消融实验结果均有力地证实了SV-RAG架构中各个创新模块的必要性。

5 结论

为应对国防安全领域的特殊需求，提出了SV-RAG架构，旨在解决标准RAG模型在处理国

防白皮书等高权威性、结构化文本时存在的语义偏差、上下文信息丢失及知识安全风险等问题。通过在知识存储、知识检索与问答生成3大核心环节进行系统性优化，并构建数据集开展实验，验证了该架构的有效性，主要结论如下。

1）在知识库构建与检索重排序阶段融合文档的层次结构元数据，是解决上下文依赖性术语理解偏差的有效途径。与忽略文本原有层次结构的传统RAG方法相比，该策略使模型能够精确识别政策术语在特定章节下的特定内涵，显著提升了问答的语境准确性，有效解决了因未利用文档结构而导致的上下文信息丢失问题。

2）所提出的双轨混合检索与交叉验证机制，为在高敏感领域安全地融合外部信息提供了可行的技术方案。研究结果表明，以权威知识库为基准对外部信息进行核查，能够有效识别并滤除潜在的错误或过时信息。说明在保证信息丰富性的同时，严格的验证流程对于维护答案的权威性与可靠性至关重要，补充了现有研究在外部知识源安全性考量方面的不足。

3）“原文引用与精准应答”生成策略是保障答案事实忠实度的有效策略。通过设定结构化的输出格式，约束了语言模型的自由生成倾向，使其输出内容严格基于权威文本。实验结果表明，在处理国防政策等措辞严谨的文本时，对生成过程施加明确约束，是规避语义偏差、确保答案与官方立场高度一致的关键机制，其在保障忠实度方面的性能优于传统的开放式生成方法。

本研究尚存一定局限，例如，关键参数如检索召回文档数量K的选择依赖于经验设定，未能进行详尽的敏感性分析以探究其对模型性能的深层影响，可能限制系统在不同配置下实现最优性能。针对上述局限，未来的研究将聚焦于参数自适应优化，探索使模型能根据查询复杂度和上下文信息动态调整检索参数的方法。此外，将尝试融合知识图谱技术，以处理超越文本匹配的深层次、多关联的复杂分析任务，推动系统从一个高效的“信息提供工具”向可信赖的“智能分析辅助系统”发展。

参考文献

[1] ANSARI S，RAUT R，SHAH B K. A review of question answering systems：approaches，challenges，and applications[J]. International Journal of Computer，2023，46(1)：25-33.

[2] KARPATHY A，LI F F. Deep visual-semantic alignments for generating image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(4)：664-676.

[3] ZHAO W X，ZHOU K，LI J Y，et al. A survey of large language models[EB/OL].(2025-03-11)[2025-08-28]. https：//arxiv.org/abs/2303.18223.

[4] ZHANG Y，LI Y F，CUI L Y，et al. Siren’s Song in the AI ocean：a survey on hallucination in large language models[EB/OL].(2023-09-05)[2025-08-28]. https：//arxiv.org/abs/2309.01219.

[5] HUANG J，CHANG K C C. A comprehensive survey on retrieval-augmented generation for large language models[J]. Neurocomputing，2024，571：127184.

[6] GAO Y F，XIONG Y，GAO X Y，et al. Retrieval-augmented generation for large language models：a survey[EB/OL].(2024-03-27)[2025-08-28]. https：//arxiv.org/abs/2312.10997.

[7] 中华人民共和国国务院新闻办公室. 新时代的中国与世界[M]. 北京：人民出版社，2019：98.

[8] ZHONG H，XIAO C，TU C，et al. A survey on large language models for legal applications[J]. ACM Transactions on Information Systems，2024，42(5)：1-15.

[9] KIEN V N，AN N T，LUONG V T，et al. Legal-eagle：a retrieval-augmented generation system for Vietnamese legal question answering[J]. Information Processing & Management，2024，61(4)：103685.

[10] XIONG Y，YANG Z，LIU X，et al. An evidence-based generative question-answering system for healthcare[J]. Journal of the American Medical Informatics Association，2024，31(2)：336-345.

[11] ZHAO W X，ZHOU K，LI J，et al. A survey of large language models[EB/OL].(2023-03-29)[2025-08-28]. https：//arxiv.org/abs/2303.18223.

[12] 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.

[13] LIU Y，YAO Y H，ZHANG X Y，et al. Trustworthy LLMs：a survey and guideline for evaluating large language models alignment[EB/OL].(2024-03-21)[2025-08-28]. https：//arxiv.org/abs/2308.05374.

[14] RAM O，LEVINE Y，DALMEDIGOS I，et al. In-context retrieval-augmented language models[EB/OL].(2023-08-01)[2025-08-28]. https：//arxiv.org/abs/2302.00083.

[15] ROBERTSON S，ZARAGOZA H. The probabilis-

tic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389

[16] ASAI A, WU Z Q, WANG Y Z, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[EB/OL].(2023-10-18)[2025-08-28]. https://arxiv.org/abs/2310.11511.

[17] YAN S Q, GU J C, ZHU Y, et al. Corrective retrieval augmented generation[EB/OL].(2024-10-07)[2025-08-28]. https://arxiv.org/abs/2401.15884.

[18] MA X B, GONG Y Y, HE P C, et al. Query rewriting for retrieval-augmented large language models[EB/OL].(2023-10-23)[2025-08-28]. https://arxiv.org/abs/2305.14283.

[19] 许山山，纪梦琪，杜鸿文. 基于双向门控循环单元神经网络和自注意力模型的航母情报关系抽取方法[J]. 火箭军工程大学学报，2023，37(4)：40-46.
XU S S, JI M Q, DU H W. Aircraft carrier intelligence relationship extraction based on bi-GRU+self-attention[J]. Journal of Rocket Force University of Engineering, 2023, 37(4): 40-46.

[20] ZHAN L N, MAO J Y, LIU Y D, et al. Dense text retrieval: a survey[J]. ACM Transactions on Information Systems, 2024, 42(4): 1-59.

[21] LIN J, NOGUEIRA R, YATES A. Pretrained transformers for text ranking: Bert and Beyond[J]. ACM Transactions on Information Systems, 2021, 39(3): 1-29.

[22] 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.

[23] LI H Y, SU Y X, CAI D, et al. A survey on retrieval-augmented text generation[EB/OL].(2022-02-13)[2025-08-28]. https://arxiv.org/abs/2202.01110.

[24] YU W H, ZHANG H M, PAN X M, et al. Chain-of-note: enhancing robustness in retrieval-augmented language models[EB/OL].(2024-10-03)[2025-08-28]. https://arxiv.org/abs/2311.09210.

[25] MENANT P, DAMERA V, GLASS M, et al. Towards verifiable text generation with evolving memory and knowledge fusion[J]. Neural Computing and Applications, 2024, 36(1): 625-638.

[26] GUO Z J, SCHLICHTKRULL M, VLACHOS A. A survey on automated fact-checking[J]. Transactions of the Association for Computational Linguistics. 2022, 34(10): 178-206.

[27] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt[EB/OL].(2023-02-21)[2025-08-28]. https://arxiv.org/abs/2108.11896.

[28] 袁 聪，姚俊萍，李晓军，等. 面向图神经网络知识图谱推理的解释方法研究综述[J]. 火箭军工程大学学报，2023，37(3)：60-70.
YUAN C, YAO J P, LI X J, et al. A survey of explanation methods for GNN-based knowledge graph reasoning[J]. Journal of Rocket Force University of Engineering, 2023, 37(3): 60-70.

[29] YANG A, LI A F, YANG B S, et al. Qwen3 technical report[EB/OL].(2025-05-14)[2025-10-10]. https://arxiv.org/abs/2505.09388.

[30] GRATTAFIORI A, DUBEY A, JAUHRI A, et al. The llama 3 herd of models[EB/OL].(2024-11-23)[2025-10-10]. https://arxiv.org/abs/2407.21783.

[31] HURST A, LERER A, GOUCHER A P. GPT-4o system card[EB/OL].(2024-10-25)[2025-10-10]. https://arxiv.org/abs/2410.21276.

Faithfulness-Enhanced Retrieval-Augmented Generation Method for National Defense Policies

LUO Yun, XU Jiajun, PENG Mingyang*, ZHU Jingdan

(College of International Studies, National University of Defense Technology, Nanjing 210039, China)

ABSTRACT: To address the issues of semantic deviation, context information loss, and knowledge security risks encountered when dealing with high-authority structured texts, such as defense white papers, a structure-aware and verification-enhanced retrieval-augmented generation (SV-RAG) was proposed based on the standard RAG. First, in the knowledge base construction stage, a "structured chunking" approach

was adopted to retain and utilize the original hierarchical structure metadata of the document while segmenting the text. Second, a dual-track hybrid retrieval mechanism was designed for the knowledge retrieval stage. For the authoritative knowledge base, a query-aware adaptive weighted reranking algorithm incorporating text similarity and metadata matching was introduced to accurately capture the context meaning. For external retrieval information, a "dual-stream cross-checking" process was implemented to verify and integrate external information based on authoritative knowledge. Finally, for the question-answer generation stage, an "original text citation and precise response" generation paradigm was designed to ensure the fidelity of the answers, using prompt engineering to force the model to cite the original core text before elaborating. To verify the effectiveness of this method, a Chinese defense security knowledge dataset containing 494 high-quality question-answer pairs was constructed. The comprehensive experimental results on this dataset show significant performance improvements of the proposed SV-RAG architecture in core indicators, such as fidelity and context accuracy, with improvements of 28.0% and 28.2%, respectively, compared with the conventional RAG baseline.

KEYWORDS: large language model; factual fidelity; knowledge retrieval; authoritative text

（编辑：于福兴）