检索增强生成中文本分块策略的实证研究

杨明可，张安达，黄煜骁，张文清，路康怡
（中国电信集团有限公司上海分公司云网运营部，上海 200120）

摘　要：大语言模型在自然语言处理领域取得了显著进展，在多个领域广泛应用。然而，这些模型在生成回答时面临幻觉问题。为解决这一问题，检索增强生成技术通过从外部知识库检索相关信息来提高模型输出的事实准确性。重点研究了检索增强生成技术中的文本分块策略，并通过实证研究评估了这些策略在电信内部文档中的表现。研究结果表明，固定大小分块和按标记分割方法在召回率和准确率上表现相对较差，适用于简单任务；递归字符和语义分块可以提升召回率，但准确率不足，适用于需要捕捉局部结构和语义信息的任务；大语言模型分块方法在召回率和准确率上均表现最佳，但其计算复杂度和资源需求较高，适用于对性能要求较高的复杂任务。为实施有效分块技术提供了最佳实践，并为优化检索增强生成以改善用户体验提供了见解。

关键词：大语言模型；检索增强生成；文本分块
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025059

0　引言

近年来，大语言模型（LLM），如OpenAI的GPT系列、Meta的Llama和百度的文心一言等，已经在人工智能领域取得了显著的成功[1]。这些模型的快速发展和广泛应用，标志着自然语言处理（NLP）技术步入了一个新的纪元。通过在大规模数据集上进行预训练，大语言模型能够捕捉到语言的深层次结构和含义，使它们能够在没有特定任务训练的情况下，通过少样本的学习，完成多种大语言模型任务[2]。这种能力使得大语言模型在多个领域中具有应用价值。当前大语言模型的应用已经扩展到了医疗[3]、教育[4]、金融[5]等关键领域，其中它们的能力被用来提高决策质量、加速研究进程以及提升服务效率。

然而，当前大语言模型持续面临的一个重大挑战是一种被称为幻觉的现象，即模型生成实际上不正确、无意义或与输入上下文脱节的输出[6]。这种现象可以以多种形式表现出来，如提供虚假信息或创造现实中不存在的完全捏造的细节。大语言模型产生幻觉的主要原因之一是在生成回答时缺乏足够的上下文。为了填补上下文不足而造成的知识空白，大语言模型可能会添加细节或捏造事实。当用户依赖这些模型获取准确信息时（如在法律或医疗环境中）会对事实产生错误的判断。为了解决这个问题，研究人员探索了各种策略，其中之一是检索增强生成。

检索增强生成通过在生成响应之前从外部知识库检索相关文档或数据来提高大语言模型输出的事实准确性[7]。这确保了大语言模型可以访问当前且可验证的信息，从而最大限度地减少生成过时或不正确内容的可能性。通过将生成过程建

收稿日期：2024−11−15；修回日期：2025−02−15

立在可靠的来源之上，检索增强生成有助于确保模型的所生成的内容实际上是正确的并且与用户的查询相关。

在检索增强生成当中文本分块是一种关键的预处理技术[8]。它将文档划分为保留语义的“块”，使文档数据更易于管理和处理。然而分块策略也带来了一些挑战。例如，选择最佳的块大小至关重要，太小的块可能缺乏足够的上下文，导致输出不精确或不相关，而太大的块可能会压垮模型的处理能力。此外，传统的固定大小分块方法可能无法考虑文本内的语义边界，从而导致意义丧失。因此，开发考虑文本结构和语义方面的自适应分块策略是最大化大语言模型的有效性的重要步骤。

为了在电信内部知识库的建设过程中，进一步提升在指定知识结构文档上的准确率，本文对不同分块策略进行了实证研究。详细探讨了检索增强生成中采用的各种分块策略，包括固定大小分块、递归分块，以及语义分块等。系统地评估这些分块策略对检索增强生成表现的影响，能够为电信内部知识库的有效构建提供理论基础和实践指导。

1　文本分块策略

在检索增强生成当中，文本分块是将大型文档拆分成较小的、易于处理的块，以便更好地与生成模型匹配，并提高检索效率和准确性。

1.1　固定大小分块

固定大小分块是检索增强生成系统中的一项基本技术，涉及将大型文本语料库分解为更小、更易于管理的片段。这种方法背后的动机是促进文本数据的有效处理。固定大小分块通过将文本划分为固定长度的块来解决此问题，从而实现更有针对性和更高效的处理。使用较小、大小统一的文本段涉及以下关键步骤。

（1）确定块大小。第一步是确定适当的块大小，通常以字符、单词或标记的数量来衡量。块大小的选择取决于各种因素。

（2）分割文本。确定块大小后，将文本分割为固定长度的块。此过程可以使用简单的字符串操作技术执行，如根据字符或单词的数量拆分文本。

（3）处理块边界。为了确保重要的上下文信息不会在块边界处丢失，可以引入相邻块之间的重叠。重叠大小通常是块大小的一小部分。这种方法有助于保持分块文本的连贯性和完整性。

（4）处理和存储块。分割后，将对块进行处理并以合适的格式存储，以便进一步分析或检索。这可能涉及其他预处理步骤，如删除停用词、词干提取或词形还原等。处理后的块可以存储在数据库中或编入索引，以便高效检索。

1.2　递归字符分块

递归字符分割是一种用于文本分块的推荐方法，也是许多检索增强生成系统默认使用的方法。这些分块方法对语料库的语义内容不敏感，而是依赖字符序列的位置将文档划分为块，直至最大指定长度。它通过一组字符来参数化分割规则，按顺序尝试使用这些字符对文本进行分割，直到文本块足够小。默认字符集包括换行符（“\n\n”“\n”）、空格（“ ”）和空字符串（“”）等。这一策略通过尽量保持段落、句子和词语的完整性，来进行切分，因为这些通常是语义关联性最强的文本片段。

文本块的大小是以字符数量来衡量的，并且可以根据需要自定义分割的字符集。例如，对于某些没有单词边界的语言，如中文、日文和泰语，默认的字符分割方式可能导致词语在分割时被拆开。为了避免这种情况，可以添加更多的标点符号作为分隔符，如 ASCII 句号“.”、全角句号“．”，以及中文或日文中的句号“。”。

1.3　按标记分割

按标记分割是一种基于语言模型的分割方

法，其目的是在处理文本时考虑模型的标记（token）限制。在分割文本为多个小块时，计算标记数量来确保不会超过语言模型的限制。由于不同的模型使用不同的分词器，因此分割时需要使用与目标语言模型相同的分词器，以保持一致性。

OpenAI创建的“tiktoken”是一种快速的字节对编码分词器[9-10]，在按标记分割方法中，用来计算文本中标记的数量。由于不同语言模型的标记方式不同，“tiktoken”可以根据目标语言模型的要求，将文本分割成符合模型标记数量限制的块。“按标记分割”依赖于分词器的计算结果来控制文本块的大小，通过这种方式，确保每个分割块都不会超出模型允许的标记数。

1.4 语义分块

语义分块是一种基于语义相关性确定文本分割点的新方法，不同于传统的固定大小分块。它通过嵌入相似度动态调整分割点，确保每个块内的句子在语义上保持密切相关性，从而形成语义连贯的文本片段。利用嵌入模型计算句子间的相似度，从而判断最佳分割位置，并可以通过设置断点百分比阈值来控制分块的细化程度。

语义分块的具体流程包括：首先，准备并加载待处理的文本数据；接着，定义语义分割器，配置相关参数（如缓冲区大小和断点百分比阈值）以确保分块结果在语义上合理；然后，检查生成的块，验证每个块内的句子是否在内容上连贯。

1.5 大语言模型分块

由于大语言模型经过大量文本数据的训练，能够以出色的流利度和连贯性捕捉和生成类似人类的语言。因此本文提出了一种新颖的文本分块方法，基于大语言模型的分块，利用这些模型的强大功能将大型文本语料库拆分为有意义且连贯的块。与依赖固定大小窗口或基于规则的方法的传统分块方法不同，基于大语言模型的分块利用大语言模型中嵌入的语言知识和理解来确定最合适的块边界。

基于大语言模型的分块的总体方法包括以下步骤。

（1）提示工程。创建一个精心设计的提示词，指示大语言模型执行文本分块。提示应提供关于所需块大小、连贯性以及与具体任务相关的任何特定标准的明确指导。

（2）迭代分块。以迭代方式将文本语料库输入到大语言模型。在每次迭代中，大语言模型都会处理文本的一部分并根据其对输入的语言结构和语义的理解生成块边界。此过程持续到整个文本被分块。

（3）块细化。应用细化步骤进一步提高块的质量。这可能涉及根据其他标准（如关键短语或主题转移的存在）合并或拆分块。

2 评估切分策略

为评估不同切分策略在电信内部文档中的表现，本文构建了一个通用的测试框架。输入知识文档，自动化构建测试数据集。

2.1 构建数据集

对于由一组知识文档组成的给定文本语料库，评估框架利用大语言模型对文档当中每一个章节生成对应的查询，以及与生成的查询相关的语料库的摘录。设计提示词，向大语言模型提供语料库中的文档、生成有关语料库的事实查询的说明，并提供与生成的问题相对应的摘录，让其自动化地构建评估数据集。

为了确保大语言模型生成的查询的有效性，框架会过滤初始生成步骤的输出。具体地，由于大语言模型偶尔会生成重复或接近重复的查询，因此框架通过嵌入每个查询并计算所有查询对的嵌入之间的余弦相似度来对生成的数据集进行重复数据删除。随后，过滤余弦相似度高于某个阈值的所有查询和相关摘录。

对于所采用的知识文档，本研究评估了来自电信行业内部的66份文档，这些文档涵盖了培训材料、规章制度以及操作指导等内容。通过纳入多种类型的知识文档，本研究能够有效地评估切分策略在实际应用场景中的效果。

2.2　评估指标

传统的信息检索基准通常侧重于检索文档的相对排名[11]。然而，在实际应用中，大型语言模型对相关信息在上下文窗口中的位置相对不敏感[12]。此外，与特定查询相关的信息可能分散在多个文档中，这使得对文档之间相对排名的评估变得模糊。

鉴于这些局限性，本文提出了从标记级别评估检索性能。利用大型语言模型生成一组查询及其相关的摘录，并从任意给定的文本语料库中筛选这些内容，随后通过检索标记的准确度和召回率来评估检索性能。这种评估采用了细粒度的方法，以评估检索的准确性和效率。在检索增强生成系统中，相关标记的存在与否比它们的相对排名更为重要。

具体而言，对于与特定语料库相关的给定查询，仅该语料库中的标记子集被视为相关。理想情况下，为了提高检索的效率和准确性，检索系统应当准确地检索整个语料库中每个查询的相关标记。然而，在实际应用中，检索增强生成系统的检索单元通常是包含相关摘录的文本块。这些文本块往往包含冗余的标记，这需要额外的计算来处理，并且可能还会包含不相关的干扰信息，从而降低检索增强生成系统的整体性能。此外，在使用重叠块的分块策略下，检索器可能会返回冗余的标记，例如，当来自相关摘录的标记因重叠而位于多个块中时。最后，给定的检索器可能无法检索到满足特定查询所需的所有必要摘录。

因此，本文从标记级别评估检索性能，其不仅可以考虑是否检索到相关摘录，还可以考虑检索到多少不相关、冗余或分散注意力的标记。对于语料库 C，则用户提出的查询记为 q。本文使用 t_e 表示所有相关摘录中的标记集合。t_r 表示所有检索到的文本块的标记集合。

本文将准确率定义如下。

$$\text{Precision}_q(C) = \frac{|t_e \cap t_r|}{|t_r|} \tag{1}$$

召回率定义如下。

$$\text{Recall}(C) = \frac{|t_e \cap t_r|}{|t_e|} \tag{2}$$

2.3　评估结果

对于所构建的测试集，不同切分方式评估结果见表1。

表1　不同切分方式评估结果

分块方法	分块长度	重叠长度	召回率	准确率
固定大小分块	400	200	81.82%	3.05%
递归字符分块	400	200	84.44%	3.22%
按标记分割	400	0	82.27%	2.37%
语义分块	不适用	0	82.66%	1.35%
大语言模型分块	不适用	0	**86.12%**	**3.77%**

比较了5种文本分块方法的精度和准确率表现，本文分析了各方法的原理及其与精度和准确率的关系。

固定大小分块方法通过将文本分割为固定长度的块，并在块之间设置重叠区域以提高召回率。该方法的召回率为81.82%，准确率为3.05%。固定大小分块的优势在于其简单性和易于实现，但其召回率相对较低，且准确率也不高，表明该方法在捕捉长距离依赖关系和复杂语义信息方面存在局限性。

递归字符分块方法通过递归地分割文本，直到达到预设的分块长度，并在块之间保留重叠区域。该方法的召回率为84.44%，准确率为3.22%。相较于固定大小分块，递归字符分块在召回率上有所提升，表明其在捕捉文本中的局部结构和上下文信息方面表现更好。然而，其准确率仍然较低，说明该方法在处理复杂语义关系时

仍存在挑战。

按标记分割方法通过识别文本中的特定标记（如标点符号、空格等）来分割文本，且不设置重叠区域。该方法的召回率为82.27%，准确率为2.37%。按标记分割的优势在于其能够较好地保留文本的结构信息，但其召回率和准确率均较低，表明该方法在处理长文本和复杂语义时表现不佳。

语义分块方法通过分析文本的语义信息来分割文本，且不设置重叠区域。该方法的召回率为82.66%，准确率为1.35%。语义分块的优势在于其能够捕捉文本的深层语义信息，但其准确率较低，表明该方法在处理歧义和多义词时存在困难。此外，语义分块的召回率也相对较低，说明其在处理长距离依赖关系时表现不佳。

大语言模型分块方法利用大语言模型来分割文本，且不设置重叠区域。该方法的召回率为86.12%，准确率为3.77%。大语言模型分块在召回率和准确率上均表现最佳，表明其在捕捉文本的语义信息和长距离依赖关系方面具有显著优势。然而，该方法的计算复杂度较高，且对计算资源的需求较大，限制了其在实际应用中的广泛使用。

3 结束语

本文深入探讨了在检索增强生成系统中，不同文本分块策略对其性能的深远影响。实证研究系统地分析了5种核心分块方法：固定大小分块、递归字符分块、按标记分割、语义分块以及大语言模型分块，并对其在电信内部知识库中的实际表现进行了全面评估。

实验结果揭示了各分块方法的优劣。并指出选择最适宜的分块方法需要基于具体的应用场景和实际需求进行权衡，以期在召回率和准确率之间找到最佳平衡点。本文的研究不仅为电信内部知识库的构建提供了坚实的理论支撑和实践指南，更为优化检索增强生成系统、提升用户体验提供了宝贵洞见。

尽管本文已取得一定研究成果，但仍有许多值得进一步探索和深化的领域。未来研究可考虑将图像、音频等多模态数据融入分块策略，以期在处理需要综合多种信息来源的复杂任务时，进一步提升检索增强生成系统的能力。同时，利用强化学习技术优化分块策略亦是一个充满前景的研究方向，设计奖励函数，训练模型在分块过程中实现策略的自动调整，从而最大化检索性能。此外，引入用户反馈机制，持续收集和分析用户的使用数据及反馈意见，迭代优化分块策略和检索增强生成系统，将有助于更精准地满足用户需求。进一步提升大语言模型的语义理解和生成质量，亦能显著增强分块策略的效果。未来可探索更为先进的预训练技术和微调方法，以期在复杂语境下提升模型的表现，进而全面提升检索增强生成系统的整体性能。

参考文献：

[1] YU J F, WANG X Z, TU S Q, et al. KoLA: carefully benchmarking world knowledge of large language models[J]. arXiv preprint, arXiv: 230609296, 2023.

[2] YANG Y H, TOMAR A. On the planning, search, andMemorization capabilities ofLarge language models[C]//Proceedings of International Conference on Intelligent Vision and Computing (ICIVC 2023). Cham: Springer Nature Switzerland, 2024: 24-38.

[3] 潘盼, 蒋硕然, 刘于红, 等. 大语言模型ChatGPT在电阻抗断层显像(EIT)肺通气诊断中的应用价值[J]. 中国急救医学, 2023, 43(10): 760-767.

[4] 杨兴睿, 马斌, 李森垚, 等. 基于大语言模型的教育文本幂等摘要方法[J]. 计算机工程, 2024, 50(7): 32-41.

[5] 陆岷峰, 高伦. 大语言模型发展现状及其在金融领域的应用研究[J]. 金融科技时代, 2023, 31(8): 32-38.

[6] HUANG L, YU W J, MA W T, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[7] FAN W Q, DING Y J, NING L B, et al. A survey on RAG meeting

LLMs: towards retrieval-augmented large language models[C]// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2024.

[8] FINARDI P, AVILA L, CASTALDONI R, et al. The chronicles of RAG: the retriever, the chunk and the generator[J]. arXiv preprint, arXiv: 240107883, 2024.

[9] MANSUROVA A, MANSUROVA A, NUGUMANOVA A. QA-RAG: exploring LLM reliance on external knowledge[J]. Big Data and Cognitive Computing, 2024, 8(9): 115.

[10] SENNRICH R. Neural machine translation of rare words with subword units[J]. arXiv preprint, arXiv: 150807909, 2015.

[11] PEREZ F, RIBEIRO I. Ignore previous prompt: Attack techniques for language models[J]. arXiv preprint, arXiv: 221109527, 2022.

[12] WANG W, ZHANG S, REN Y, et al. Needle In A Multimodal Haystack [J]. arXiv preprint, arXiv: 240607230, 2024.

[作者简介]

杨明可（1998– ），男，现就职于中国电信股份有限公司上海分公司云网运营部，主要从事知识库原型算法研究相关工作。

张安达（1995– ），男，现就职于中国电信股份有限公司上海分公司云网运营部，主要从事大模型应用和RAG相关工作。

黄煜骁（1993– ），男，现就职于中国电信股份有限公司上海分公司云网运营部，主要从事AI相关基础平台和应用能力建设规划和总体架构设计工作。

张文清（1999– ），女，现就职于中国电信股份有限公司上海分公司云网运营部，主要从事异构文档智能解析相关研究工作。

路康怡（1999– ），女，现就职于中国电信股份有限公司上海分公司云网运营部，主要从事知识库原型算法研究相关工作。