Chunk-Link: Context-aware chunk completion*

Kenichirou Narita[1,*], Satoshi Munakata[1]

[1]Fujitsu Ltd., 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa 211-8588, Japan

Abstract

In context engineering, Retrieval Augmented Generation (RAG) is an essential technology for improving the reliability of generative AI. However, conventional Top-K vector search methods often face challenges, particularly when enumerated content (such as bulleted lists or other list-based information) spans multiple chunks. This often leads to incomplete retrieval and inaccurate answers due to missing contextual fragments. To address this, we propose a novel method, "Chunk-Link." It identifies enumeration relationships across chunks to comprehensively extract all necessary contextual fragments. Our evaluation shows that Chunk-Link significantly improves chunk extraction accuracy. This, in turn, leads to a higher overall answer quality and enables more precise information retrieval and answer generation.

Keywords

Large Language Model, Retrieval-Augmented Generation, Context Engineering, Multi-Chunk Information Handling,

1. Introduction

In commercial question-answering (QA) applications, it is essential to accurately extract necessary and sufficient context from specialized and up-to-date documents and provide it to Large Language Models (LLMs).While Top-K vector search is the most popular method for retrieving relevant chunks in Retrieval Augmented Generation (RAG), it often fails to retrieve all correct chunks when enumerated information (such as bullet points or lists) or complex concepts span across multiple chunks. This leads to incomplete answers. This challenge, related to Multi-Chunk Information Handling, is frequently observed in real-world documents like guidelines and manuals.

In this paper, we propose "Chunk-Link," a novel method that enables LLMs to identify enumeration relationships spanning across chunks, thereby complementing conventional vector search. By leveraging inter-chunk relationships to supplement chunks missed by traditional vector search, Chunk-Link provides LLMs with sufficient context.

2. Related work

Enhancing RAG performance, especially when information is distributed across multiple chunks, has become a key research area in recent years. This chapter focuses on the challenge of "Multi-Chunk Information Handling" and reviews relevant prior research.

2.1. Text Chunking Strategy

Common chunking strategies include fixed-length chunking, recursive chunking, and semantic chunking. Semantic chunking, in particular, analyzes the content of text and generates chunks based on

RAGE-KG 2025: The Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2025, November 2–6, 2025, Nara, Japan

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

k.narita@fujitsu.com (K. Narita); munakata.satosi@fujitsu.com (S. Munakata)

0000-0003-3974-5918 (K. Narita)

semantic boundaries and contextual continuity. However, it may require adjustments to account for domain-specific knowledge and structure, and it can be challenging to balance the granularity of chunk division, potentially failing to fully capture the structure and semantic coherence of a document [1].

2.2. Multi-Chunk Information Handling

To address cases where complex questions cannot be answered with a single chunk, Kwon et al. proposed an approach that decomposes questions into multiple subqueries and searches for and integrates chunks corresponding to each subquery [2]. Approaches that dynamically change chunk sizes have also been proposed. Techniques such as Parent Document Retrieval perform searches using small chunks and provide larger context blocks to the LLM based on the results to prevent context loss. Zhong et al. dynamically determine the optimal granularity (chunk size or level of grouping) of the knowledge database based on the input query and provide more appropriate context to the LLM [3]. Hei et al. propose a two-stage search process to extract chunks that are not directly related to the query. After extracting chunks highly related to the query, they combine documents related to the query with the extracted chunks to further search for "dynamically related documents" that may appear unrelated but are essential for the answer [4].

These studies primarily focus on the relationship between input queries and chunks. However, there is still insufficient focus on the relationship between chunks themselves, particularly on explicitly distinguishing the "enumeration relationship" between chunks and comprehensively extracting necessary chunks based on such relationships.

3. Problems and Challenges

In documents such as guidelines and manuals, product lists and operating procedures are commonly summarized in a list format. When querying an LLM about such list-based information, traditional vector search often omits parts of the enumerated content from the generated response. **Fig.1** illustrates a common scenario where enumerated content spans multiple chunks (here, each chunk corresponds to one page). The text relevant to the enumeration, as shown in the bounding box, is distributed across several pages. When extracting chunks from such a document using the following query, conventional vector search cannot correctly extract the necessary chunks:

What specific expansion efforts will NIST undertake to enable breakthroughs in measurement, standards, and process capabilities for the fabrication of next-generation semiconductors?

This failure occurs because conventional search primarily focuses on the direct semantic relationship between queries and individual chunks, and thus cannot detect the complex structural relationships between chunks that constitute a complete enumeration. In order to provide adequate answers when list-based information is required and spans multiple chunks, it is essential to proactively identify these inter-chunk relationships. This allows for chunk retrieval that considers not only query-chunk relevance but also the logical connections between chunks.

4. Chunk-Link

To address this issue, we propose "Chunk-Link." This method focuses on inter-chunk enumeration relationships and integrates them into conventional chunk search. Chunk-Link maintains these enumeration relationships as explicit inter-chunk relations, and by complementing conventional query-chunk vector search, it expands the chunk search scope (**Fig.2**).

4.1. Contextual Enumeration Analysis

In this proposal, we focus on the context of the text when extracting enumerations. In documents such as books, reports, and presentations, enumeration information is written after some kind of introduction. We define sentences that promote reader understanding and clarify the existence of enumerations

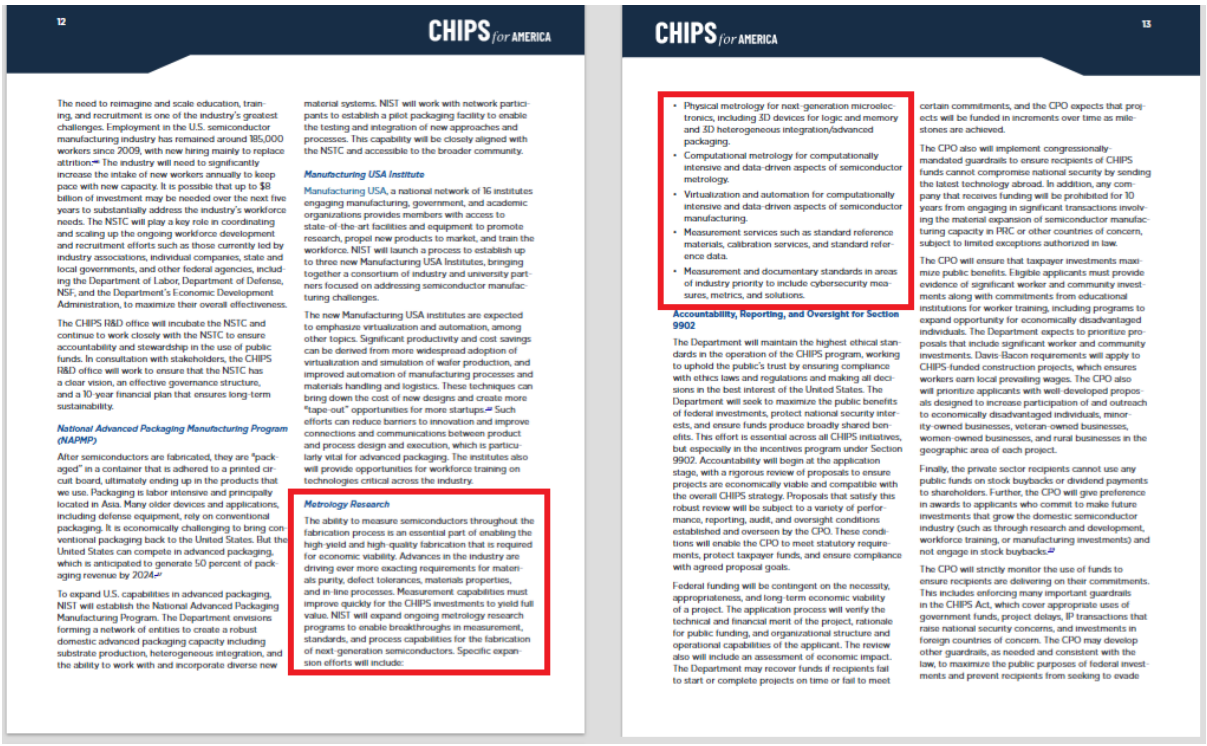

Figure 1: America-Strategy (Sept 6, 2022) page.12-13 Photograph by Manufacturing.gov. [Public domain], via A Strategy for the CHIPS for America Fund. (https://www.manufacturing.gov/reports).

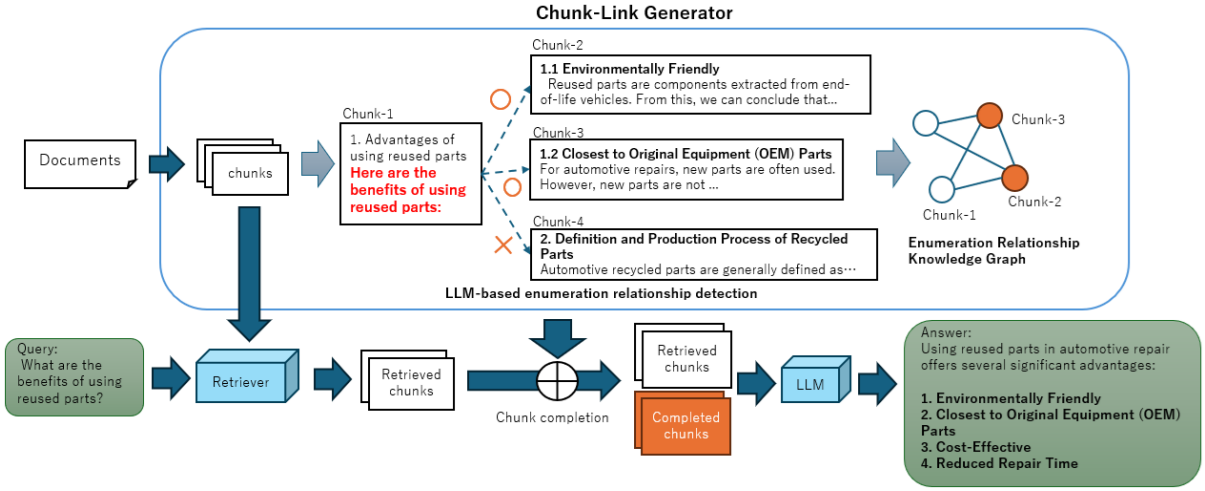

Figure 2: Chunk-Link Overview

presented later as "enumeration definition sentences." Enumeration definition sentences are not limited to sentences in the main text; chapter and section titles can also serve this role. The following are some examples.

- The following
- Next
- The concrete solutions are three: A, B, and C.
- 2-1. Advantages of using manufacturer parts

  (announces that specific benefits will be listed or explained in parallel afterwards)

Such enumerative definitions suggest that there are enumerative sentences that follow in the context, and LLM can extract enumerative information related to them from enumerative definitions. In this proposal, we extract enumerative relationships between chunks through the following steps.

1. Prepare the target chunk (any one of the chunks used in RAG) and the subsequent chunks (chunks following the target chunk in the same document).
2. Extract the enumeration definition sentences from the target chunk(using LLM). (The prompt used is provided in Appendix A).
3. Extract the enumeration sentences related to the enumeration definition sentences from the subsequent chunks(using LLM).

By focusing on enumeration preamble sentences within the context and extracting enumeration relationships, a knowledge graph representing inter-chunk enumeration relationships can be designed.

4.2. Complementing RAG with Chunk-Link

Chunk-Link operates as a complementary function for LLM-based RAG technology. In the knowledge graph generated by Chunk-Link, entities represent chunks, and the edges between them denote enumeration relationships. By correlating the chunks obtained during RAG's retrieval phase with this knowledge graph, it becomes possible to supplement the chunks retrieved by conventional RAG with those having enumeration relationships.

This ensures that LLMs receive all necessary and sufficient chunks related to enumerations.

5. Experiment

We evaluate the contribution to LLM's answer accuracy for enumeration questions. The LLMs used for extracting enumeration relationships with Chunk-Link are gpt-4o-mini and gemma-3-4b-it. OpenAI API was used for generating the embedding vectors for conventional RAG. The LLM model used to generate the final answer from the supplemented chunks and query was gpt-4o-mini.

The experiment is conducted for each setting (Chunk-Link enabled/disabled) on the same dataset for three independent runs, and their average values are reported. This ensures the stability and reliability of the results. The statistical significance of the differences in these average values is also evaluated, and their reliability is discussed.

5.1. Dataset

To evaluate the answer accuracy of LLMs focused on enumeration problems, the authors created a dataset that compiles enumeration problems. This dataset was created from text data extracted from 29 PDF files in Japanese and English, and consists of 186 queries. The documents were selected from the RAG evaluation dataset allganize_rag_evaluation_dataset_ja[5] and public documents (such as guidelines and reports published by the National Institute of Standards and Technology (NIST) and the UK Government). An example of the dataset is shown in **Table 1**.

When the created dataset was fed to an LLM (gpt-4o-mini), the accuracy using RAG was 0.504, and when the chunks were intentionally specified as only correct chunks, the accuracy was 0.856. These results indicate that this dataset can achieve high accuracy if the correct chunks can be extracted, but the queries cannot maintain that accuracy using conventional vector search.

5.2. Experimental Setup

Compare cases where RAG is supplemented with Chunk-Link and cases where it is not. The experimental parameters are shown in **Table 2**.

Table 1
Sample Queries for Dataset

query	points	relevant_pages
Which elements work together in the Framework Core to support the management of cybersecurity risk?	* Functions * Categories * Subcategories * Informative References	[13,14]
Which general types of control gates are included in the security considerations for the implementation/assessment phase?	* System Test Readiness Review * C&A Review * Final Project * Status and Financial Review...	[11,12]

Table 2
Experimental Parameters

Items	Values	Comments
top_k	5	Top-K vector search
chunk range	5	Search subsequent chunk ranges for enumeration relations

6. Results

The evaluation results are shown in **Table 3**. The effectiveness of Chunk-Link was evaluated from two perspectives: chunk extraction accuracy and query response accuracy. The chunk search accuracy was evaluated using recall, precision, and F1. When using Chunk-Link, both models showed an increase of approximately 15% in recall and a decrease of approximately 5% in precision, indicating that the enumeration relationships between chunks were extracted with high accuracy. As recall improved, the response accuracy also increased in both models.

Table 3
Comparison of Model Performance with and without Chunk-Link Usage, including Performance Changes

Model	Chunk-Link Status	Recall	Precision	F1	Accuracy
gemma	Disabled	0.601	0.209	0.310	0.499
gemma	Enabled	0.740	0.164	0.268	0.522
Change (enabled vs. disabled)		**+0.139**	**-0.045**	**-0.042**	**+0.023**
gpt-4o-mini	Disabled	0.601	0.209	0.310	0.504
gpt-4o-mini	Enabled	0.750	0.188	0.301	0.535
Change (enabled vs. disabled)		**+0.149**	**-0.021**	**-0.009**	**+0.031**

The statistical significance of the measurement results was evaluated. Since the Recall differences of the chunks did not follow a normal distribution, the Wilcoxon signed-rank test was adopted for the analysis. A significance level of α=0.05 was set. As a result, the Recall differences were below the significance level, indicating statistical significance. On the other hand, although the statistical significance of Accuracy was also evaluated using McNemar's test, the P-value was 0.190, and no statistically significant difference was observed.

Next, we analyze the impact of Chunk-Link on the level of detail in model responses through specific success examples. We introduce how the example shown in **Fig.1** was improved using Chunk-Link. The response results with and without Chunk-Link are shown in **Table 4**. As can be seen from this table, the predicted answers without Chunk-Link are limited to general explanations, while those with Chunk-Link are confirmed to include highly detailed and specific information. In particular, the latter answers list specific expansion efforts such as "physical measurements for next-generation microelec-

tronics," which is the result of Chunk-Link referencing more extensive and relevant information chunks. This level of detail is further supported by the increase in 'chunk num' from 5 to 6 and the significant improvement in 'recall' from 0.5 to 1.0. This enables more comprehensive and accurate information provision in response to user inquiries.

Table 4
Examples of improved prediction accuracy using Chunk-Link

Items	Chunk-Link Disabled	Chunk-Link Enabled
predict	NIST will expand ongoing metrology research programs to enable breakthroughs in measurement, standards, and process capabilities for the fabrication of next*generation semiconductors.	NIST will expand ongoing metrology research programs to enable breakthroughs in measurement, standards, and process capabilities for the fabrication of next*generation microelectronics. * Specific expansion efforts will include physical metrology for next*generation microelectronics, computational metrology for computationally intensive and data*driven aspects of semiconductor metrology, virtualization and automation for semiconductor manufacturing, measurement services such as standard reference materials, calibration services, and standard reference data, and measurement and documentary standards in areas of industry priority including cybersecurity measures, metrics, and solutions.
chunk num	5	6
recall	0.5	1.0
precision	0.2	0.333

7. Limitations

Our evaluation relies on a small dataset of enumeration problems from specific document types (e.g., guidelines, manuals), limiting the reliability of answer accuracy and the statistical significance of validation results. Furthermore, as Chunk-Link complements conventional vector search, its fundamental retrieval accuracy is inherently dependent on query-vector similarity, potentially imposing a ceiling on final answer accuracy.

While Chunk-Link improves Recall, Precision decreases due to irrelevant enumeration relationships, hindering overall answer accuracy. Filtering extracted enumeration relationships is essential.

8. Conclusion

In this paper, we propose "Chunk-Link" to address the issue that conventional chunk search from embedded vectors struggles to extract inter-chunk relationships. Chunk-Link extracts enumeration relationships between chunks and complements vector search. We created a dataset to measure the answer accuracy for enumeration questions and demonstrated that by complementing RAG with Chunk-Link, the recall of chunks improves even when enumerated content is divided across chunks, enabling more comprehensive and detailed answers.

For future work, we will generalize to other structural relationships, such as contextual, referential, and causal relationships, to expand Chunk-Link's applicability and versatility.

A. Prompts for LLM

This appendix provides the prompts used for extracting enumerated definitions from the context for our research. The prompts listed are excerpts of key parts, with output formats and examples omitted.

Prompt to extract enumeration definition sentences

```
You are a text analysis assistant. Following the specifications below, accurately
and comprehensively extract all "Preamble Sentences" from the input text.

### 1. Definition and Role of "Preamble Sentence"
**A "Preamble Sentence"** refers to a description that meets any of the following
conditions and **explicitly announces that multiple specific pieces of information
or elements will be enumerated immediately after it, either in a list format
(e.g., bullet points or numbered lists) or in a clearly parallel structure.**
1.  **Introductory/Announcing Type Description:**
    A short expression that **explicitly states within the sentence or at its end**
    that multiple elements will follow immediately after it, either as **bullet
    points (with leading symbols or numbers)**, or in a **clearly identifiable
    parallel structure (e.g., "A, B, and C" or "first A, then B, finally C,").**
    * **Characteristics:**
        This refers to a sentence that, within the text or at its end, is followed
     by multiple elements (either as bullet points or enumerated in subsequent sentences).
   * **Key Phrase Examples (detected in conjunction with clear subsequent enumeration):**
        * "The following ~"
        * "Next ~"
        * "...are as follows."
        * "In ..., three points are important: A, B, and C."
        * Note: Expressions like "is" alone or "regarding ~" that merely indicate
        a continuation of explanation are generally not considered key phrases.
2.  **Chapter/Section Title:**
    A title placed at the beginning of a specific chapter or section in a document
    structure, which **summarizes the specific content to be covered within that
    entire chapter/section, and functions as an introduction or announcement for the
    subsequent specific descriptions or subsections.**
        * "2-1. Advantages of using manufacturer parts"
        * "Chapter 3: Customer Experience Improvement Strategies"

### 2. Extraction Procedure
1.  Parse the input text and segment it into sentences as much as possible, including
   titles, body text, and bullet points. Then, assign a sequential number starting from
    1 to each segmented sentence.
2.  For each sentence/chapter/section title, determine if it is a "Preamble Sentence"
   based strictly on the definition and extraction conditions. Special emphasis should
    be placed on "whether multiple elements are clearly enumerated in bullet points or
    a parallel structure immediately after it."
3.  Format the output according to the output specification and output it in JSON format.

Following these rules, extract all preamble sentences from the document comprehensively.
### Input Text
```


Acknowledgments

This project would not have been possible without the immense cooperation and support of all project members and colleagues, to whom I extend my heartfelt gratitude. We would especially like to thank Ms.Moriyama for her valuable assistance in creating the dataset.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini and Copilot in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Multimodal, Semantic chunking for rag: Better context, better results, Multimodal.dev (2024). https://www.multimodal.dev/post/semantic-chunking-for-rag.

[2] M. Kwon, J. Bang, S. Hwang, J. Jang, W. Lee, A dynamic-selection-based, retrieval-augmented generation framework: Enhancing multi-document question-answering for commercial applications, Electronics 14 (2025) 659. doi:10.3390/electronics14040659.

[3] Z. Zhong, H. Liu, X. Cui, X. Zhang, Z. Qin, Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation, in: Proceedings of the 31st International Conference on Computational Linguistics, Association for Computational Linguistics, 2025, pp. 5756–5774. URL: https://aclanthology.org/2025.coling-main.384.

[4] Z. Hei, W. Liu, W. Ou, J. Qiao, J. Jiao, G. Song, T. Tian, Y. Lin, Dr-rag: Applying dynamic document relevance to retrieval-augmented generation for question-answering, arXiv preprint arXiv:2406.07348 (2024).

[5] Allganize, RAG Evaluation Dataset JA, 2024.