A New HOPE: Domain-agnostic Automatic Evaluation of Text Chunking

Henrik Brådland
Centre for Artificial Intelligence Research, University of Agder
Kristiansand, Agder, Norway
Norkart AS
Oslo, Norway
henrik.bradland@uia.no

Morten Goodwin
Centre for Artificial Intelligence Research, University of Agder
Kristiansand, Agder, Norway

Per-Arne Andersen
Centre for Artificial Intelligence Research, University of Agder
Kristiansand, Agder, Norway

Alexander S. Nossum
Norkart AS
Oslo, Norway

Aditya Gupta
Centre for Artificial Intelligence Research, University of Agder
Kristiansand, Agder, Norway

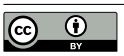
Abstract

Document chunking fundamentally impacts Retrieval-Augmented Generation (RAG) by determining how source materials are segmented before indexing. Despite evidence that Large Language Models (LLMs) are sensitive to the layout and structure of retrieved data, there is currently no framework to analyze the impact of different chunking methods. In this paper, we introduce a novel methodology that defines essential characteristics of the chunking process at three levels: intrinsic passage properties, extrinsic passage properties, and passages-document coherence. We propose HOPE (Holistic Passage Evaluation), a domain-agnostic, automatic evaluation metric that quantifies and aggregates these characteristics. Our empirical evaluations across seven domains demonstrate that the HOPE metric correlates significantly ($\rho > 0.13$) with various RAG performance indicators, revealing contrasts between the importance of extrinsic and intrinsic properties of passages. Semantic independence between passages proves essential for system performance with a performance gain of up to 56.2% in factual correctness and 21.1% in answer correctness. On the contrary, traditional assumptions about maintaining concept unity within passages show minimal impact. These findings provide actionable insights for optimizing chunking strategies, thus improving RAG system design to produce more factually correct responses.

Keywords

Document Chunking, Passage Evaluation, Retrieval-Augmented Generation, Text Embedding, Natural Language Processing

ACM Reference Format:
Henrik Brådland, Morten Goodwin, Per-Arne Andersen, Alexander S. Nossum, and Aditya Gupta. 2025. A New HOPE: Domain-agnostic Automatic Evaluation of Text Chunking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3726302.3729882



This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGIR '25, July 13–18, 2025, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3729882

1 Introduction

Document chunking, the process of dividing texts into coherent passages, plays a crucial role in RAG architectures [35]. Recent studies demonstrate that chunking strategies impact downstream task performance [34, 37, 1] and that LLMs exhibit sensitivity to passage format and structure [3, 32]. However, the field lacks formal definitions and evaluation methodologies that directly assess chunking quality independent of downstream applications. Current approaches primarily evaluate chunking through end-task performance, leaving unexplored fundamental questions about passage quality characteristics. To address this gap, this paper has the following main contributions:

- A new, more open definition of chunking.
- Three formalized principles for chunking.
- The HOPE metric, a holistic evaluation of passage quality.
- Experiments showcasing the impact of HOPE.

1.1 Paper Outline

In section 1 the importance of chunking is stated, a definition is presented and principles for chunking are formulated. Section 2 presents related work, while section 3 introduces the methodology and the main contribution: The HOPE metric. Section 4 describes the experimental setup and presents the empirical results, which are then discussed in section 5. The conclusion is given in section 6.

1.2 Defining Chunking

Chunking traditionally refers to dividing documents into smaller text segments for Information Retrieval (IR) tasks [31]. However, with the advent of Large Language Models (LLMs) and their enhanced semantic understanding capabilities, we propose broadening this definition to encompass more sophisticated transformations of documents into passages.

Different document types present distinct chunking challenges. Academic literature typically contains long sentences with a hierarchical structure, while computer log files often consist of brief, independent entries. Moreover, documents frequently contain mixed formats, including tables, lists, and figures, which traditional chunking methods handle poorly. These variations motivate the need for a more flexible framework.

We propose a formal definition of chunking to accommodate this broader perspective:

Definition 1 (Chunking)**.** A transformation $T : D \rightarrow \mathcal{P}$ that maps a document D to a set of passages $\mathcal{P} = \{p_1, p_2, \dots\}$ such that: 1) The semantic information of D is preserved in $\mathcal{P}$. 2) The transformation may be reversible but is not required to be.

Although recent advances in RAG, such as GraphRAG [6] and KAG [15], demonstrate the potential to integrate knowledge graph construction with chunking, we focus our evaluation methodology specifically on the passage creation process to maintain clear scope.

1.3 Principals of Chunking

We intend to formalise principles for chunking based on the workings of embedding models and RAG systems. These principles then serve as a starting point for the evaluation of passages. As pointed out by Wu et. al. [31], each passage should convey a single core concept. When multiple concepts are present in a passage, the resulting vector representation can be noisy. Thus, the underlying concepts are not well represented in the embedding space. The first principle of chunking is therefore:

Principle 1. *Passages should convey one core concept.*

Current RAG architectures evaluate semantic similarity primarily through query-passage relationships without comprehensively accounting for interpassage semantic dependencies. While an initial retrieved passage may contain query-relevant information, its complete interpretation often depends on contextual information stored in non-retrieved passages, leading to responses generated on incomplete information. Recent work by Zhong et al. [37] addresses this limitation through graph-based architectures that capture passage-passage semantic relations. However, this approach introduces additional computational complexity. A more fundamental solution lies in ensuring the semantic independence of the passage during the chunking process, making the semantic information an intrinsic property of the passages.

Principle 2. *Passages should be semantically independent.*

Building upon the definitions of chunking (Definition 1), the third principle addresses the preservation of semantic information during the chunking process. While semantic preservation may seem straightforward, its implications for system performance are profound. Information loss at the chunking stage affects all later stages of RAG pipelines, including indexing, retrieval, and generation. Potentially compromising the effectiveness of downstream tasks and the reliability of generated responses. Therefore, we propose our third principle:

Principle 3. *The passage set should convey all the semantic information found in the source document.*

The provided principles are not absolute, and on several occasions, it will not be possible to fulfill all of them due to the structure of the underlying data. Occasionally, the principles will be conflicting such as having content with significant contradictions and exceptions, or with complex information relations, will struggle to fulfill the provided principles. However, any universal chunking method should strive to meet the principles, and the principles should form the basis for a universal chunking metric.

2 Related Work

This section provides an overview of previous work to support the need for a chunking evaluation methodology and background knowledge on chunking methods and NLP metrics.

2.1 LLM Sensitivity to Context

The effectiveness of LLM-based systems depends significantly on how retrieved information is structured and presented in the prompt. Cuconasu & Trappolini [3] demonstrated that distracting passages, while topically related, can reduce answer accuracy from 56.42% to 17.95% for a Llama2 model [25]. Their work established that passage ordering within prompts is crucial, with information proximity to the question strongly correlating with its influence on the response. Wu & Xie [32] further quantified this sensitivity, showing that exposure to irrelevant data can flip correct responses to incorrect ones in 15% of cases event for strong models like GPT-4 [21]. These findings underscore the importance of sophisticated chunking strategies that minimize noise while preserving semantic coherence in the retrieved passages.

2.2 Related Fields

Chunking, as defined in this article (Definition 1), shares characteristics with other Natural Language Generation (NLG) tasks, particularly summarization and Machine Translation (MT). Like summarization, chunking must preserve semantic independence between segments (principle 2), while like MT, it must maintain contextual meaning (principle 3). Traditional evaluation metrics for these NLG tasks, such as BLEU [22] and ROUGE [2], rely on lexical overlap between generated outputs and reference responses. However, these metrics inadequately capture semantic relationships crucial for chunking evaluation. While regression-based metrics like BLEURT and RUSE leverage pre-trained language models for improved semantic understanding, they require additional annotated datasets and face challenges with annotator bias [13].

2.3 Chunking Methods

There are several developed chunking methods of various complexity [17], which can be roughly categorized into general-purpose methods and data-specific methods. This section focuses on general-purpose methods that maintain broad applicability across different types of documents and retrieval tasks.

Fixed-size Chunking. The most fundamental approach, fixed-size chunking divides text into segments of predetermined character length. This method uses two hyperparameters: *passage size* and *overlap size* . While effective for uniform text, this approach shows limitations with structured documents.

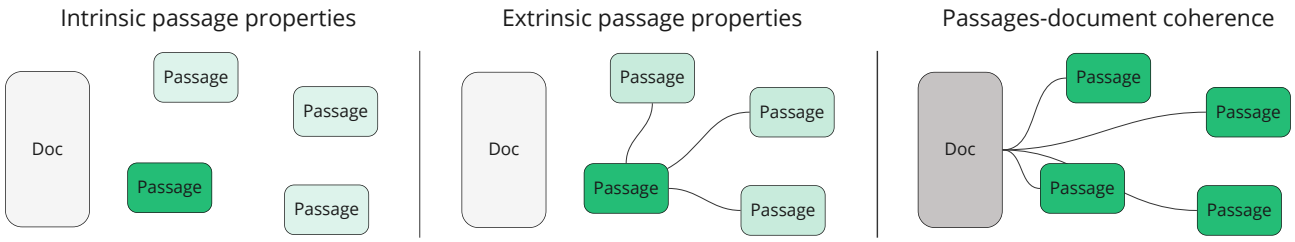


Figure 1: The three levels of holistic chunking evaluation. Left: The intrinsic passage properties are concerned with the information isolated within a passage. Middle: Extrinsic passage properties relate to how passages impact each other. Right: Passages-document coherence focuses on how well the passage set represents the original document.

Recursive Character Chunking. This method splits text recursively on delimiter patterns, following a priority order. Commonly: double newlines, single newlines, periods, and whitespace. The process continues until all passages are below a specified maximum size, making it particularly suitable for documents with clear hierarchical organization.

Semantic Chunking. Semantic chunking employs embedding models to create semantically cohesive segments. The process begins with sentence-level segmentation, followed by merging segments based on their embedding similarity. Although this approach better preserves semantic relationships, its performance depends significantly on the chosen embedding model and does not take into account semantic relations that are further apart.

3 Method

Building on the definition of established principles on chunking we introduced in this paper, we propose a three-level evaluation approach that examines passages at intrinsic, extrinsic, and set levels, as illustrated in Figure 1. This approach forms the foundation for Holistic Passage Evaluation (HOPE), a domain-agnostic metric that requires no annotated data or human intervention. The evaluation methodology directly maps to the three fundamental principles of chunking. At the intrinsic level, we evaluate concept unity (principle 1), which measures how well individual passages maintain coherent information boundaries. The extrinsic level addresses external semantic dependence (principle 2), quantifying the independence between passages. Finally, at the set level, we assess collective information preservation (principle 3), which measures how well the complete set of passages preserves the document's original information. HOPE aggregates these three properties through arithmetic mean to produce an overall score, the HOPE metric. This design enables detailed analysis of a chunking method's strengths and weaknesses across all essential dimensions. The following subsections detail our approach to quantifying each property and their integration into the unified HOPE metric.

This part of the method utilizes LLMs as a part of several algorithms. We therefore operate with the notation

$$LLM(a_1, a_2, \ldots, a_n) \rightarrow b$$

, where $a_1, a_2, \ldots, a_n$ are text elements that are part of the input prompt, and b is the LLMs text response.

3.1 Semantic Comparison

An essential component of HOPE is to compare the semantics of two or more texts. Most approaches in the literature on semantic comparison typically fall into two categories: dedicated pairwise text similarity models, such as BERT [4], and embedding-based methods that measure distances between dense vector representations [9]. HOPE implements an embedding-based approach, which offers significant computational advantages. While pairwise comparison models require processing each text pair combination, embedding-based methods need to process each text segment only once, resulting in $O(n)$ time complexity versus $O(n^2)$ for pairwise comparisons. Our implementation utilizes Qwen 2.5 [14], a state-of-the-art embedding model. The semantic comparison process consists of two steps, as shown in equation 1. First, an embedding model $\sigma(\cdot)$ encodes input text into dense N-dimensional vectors that capture semantic properties [23]. Then, the semantic similarity between two text snippets x_1 and x_2 is computed using cosine similarity $\theta(\cdot)$, which measures the angle between their vector representations $\vec{x}_1$ and $\vec{x}_2$.

$$\begin{aligned} \sigma : x &\rightarrow \vec{x} \in \mathbb{R}^N \\ \theta(x_1, x_2) &= \frac{\sigma(x_1) \cdot \sigma(x_2)}{||\sigma(x_1)|| \cdot ||\sigma(x_2)||} \in [-1, 1] \end{aligned} \tag{1}$$

While this approach efficiently captures semantic relationships, it is important to note that embedding-based methods may occasionally miss nuanced semantic differences that dedicated similarity models can detect.

3.2 Concept Unity

Derived from principle 1 of chunking, "concept unity" states that a passage should have one distinct semantic meaning. While evaluating semantics presents challenges due to the complexity of natural language, LLMs offer a robust solution for assessing concept unity in text segments, demonstrating strong semantic understanding while providing consistency and scalability [36].

Inspired by claim decomposition [28, 8], we propose using an LLM with a non-zero temperature parameter to generate a set of statements $\mathcal{S} = \{s_1, s_2, \ldots\}$ related to a selected passage $p^* \in \mathcal{P}$. The temperature parameter introduces controlled variability in statement generation, enhancing concept coverage. If the passage contains a single concept, all generated statements should exhibit

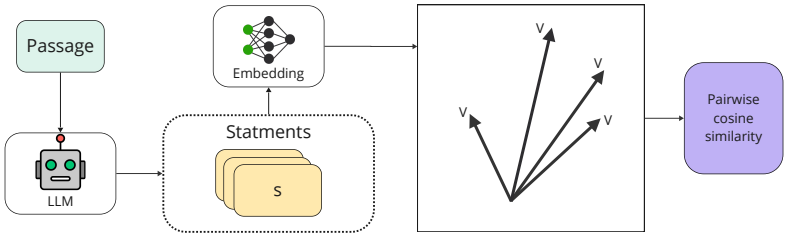


Figure 2: Calculating the concept unity $\bar{\zeta}_{con}$: The passage in green is forwarded to an LLM to produce a set of statements $\mathcal{S}$. The statements are then transformed into vector representations using an embedding model. The pairwise cosine similarity between the vectors is calculated.

high semantic similarity, which we measure using the cosine similarity of vector embeddings as shown in equation 2 and illustrated in Figure 2.

$$\begin{aligned} LLM(p^*) \rightarrow & \mathcal{S} = \{s_1, s_2, \dots\} \\ \bar{\zeta}_{con} = & \frac{1}{|\mathcal{S}|^2} \sum_{s_j \in \mathcal{S}} \sum_{s_i \in \mathcal{S}} \theta(s_i, s_j) \end{aligned} \tag{2}$$

The Concept Unity ζ_{con} is bounded by the cosine similarity interval $[-1, 1]$. A value close to 1 indicates a homogeneous set of concepts, optimal for principle 1 compliance. Negative values indicate opposing concepts, while values near 0 indicate distantly related concepts, both undesirable. We set negative values to 0, bounding $\bar{\zeta}_{con}$ to $[0, 1]$. An aggregated value ζ_{con} for the entire collection of passages $\mathcal{P}$ is computed as shown in equation 6.

$$\zeta_{con} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \bar{\zeta}_{con}(p, \mathcal{P}) \tag{3}$$

To illustrate the practical application of concept unity measurement, consider the following examples demonstrating low and high concept unity:

Low Concept Unity

```
Joe walked to the store to buy some veggies for
dinner, but the fact that the boss had purchased
a pink Mercedes was something he did not understand.
```

High Concept Unity

```
Joe walked to the store to buy some veggies for
dinner, but he could not decide between green,
red or yellow bell pepper.
```

The first example contains two concepts: Joe at the grocery store, and Joe wondering why his boss had bought a pink Mercedes. These two concepts are not related, and by the first principle of chunking should not be in the same passage, thus a low Concept Unity. The second example only contains information about Joe at the store, hence only a single concept, and high Concept Unity.

3.3 Semantic Independence

According to our definition of principle 2 of chunking, passages should be semantically independent. This independence implies that a passage's interpretation should remain consistent regardless of other passages present in the context. When an LLM performs open-book Q&A, the interpretation of any given passage should remain stable, unaffected by the presence of other passages.

We propose a method for evaluating the semantic independence of a passage p^* relative to the remaining passages $\mathcal{P}$. First, an LLM with a non-zero temperature parameter generates a set of questions $\mathcal{Q} = q_1, q_2, \dots$ based on the information from p^*. By design, the information in p^* should be sufficient to answer these questions. When passages are semantically independent, the LLM's response a^* should remain consistent even when a subset $\mathcal{P}_q \subset \mathcal{P}$ is present, as shown in equation 4.

$$\begin{aligned} LLM(q, p^*) &\rightarrow a^* \\ LLM(q, p^*, \mathcal{P}_q) &\rightarrow a \\ a^* &\simeq a \end{aligned} \tag{4}$$

The subset $\mathcal{P}_q$ comprises the top-k passages from $\mathcal{P}$ that maximize semantic similarity to the question q, where k is set to 3 in our implementation. The two LLM configurations from equation 4 each generate a set of answers $\mathcal{A} = \{a_1, a_2, \dots\}$ and $\mathcal{A}^* = \{a_1^*, a_2^*, \dots\}$ to the questions in $\mathcal{Q}$. These answers are compared pairwise using cosine similarity $\theta(a^*, a)$, as shown in equation 5.

$$\begin{aligned} LLM(q, p^*), \forall q \in \mathcal{Q} &\rightarrow \mathcal{A}^* = \{a_1^*, a_2^*, \dots\} \\ LLM(q, p^*, C_q), \forall q \in \mathcal{Q} &\rightarrow \mathcal{A} = \{a_1, a_2, \dots\} \\ \bar{\zeta}_{sem} &= \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}, a^* \in \mathcal{A}^*} \theta(a^*, a) \end{aligned} \tag{5}$$

$\bar{\zeta}_{sem}$ is bounded by the cosine similarity interval $[-1, 1]$. A value of 1 indicates complete semantic independence, while values near 0 suggest strong semantic dependence. Values approaching -1 indicate inverted semantics, which can potentially mislead LLMs [32]. Since both semantic alteration and inversion are undesirable, we cap $\bar{\zeta}_{sem}$ to the interval $[0, 1]$ by setting negative values to zero. Figure 3 illustrates the calculation process of $\bar{\zeta}_{sem}$. An aggregated value ζ_{sem} for the entire collection $\mathcal{P}$ is computed as the average of individual passage scores, as shown in equation 6.

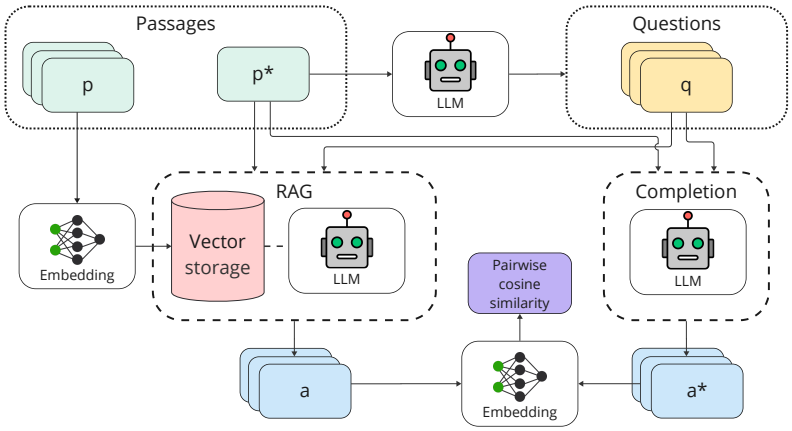


Figure 3: Calculation of the semantic independence $\bar{\zeta}_{sem}$: A selected passage p^* is used to construct a set of questions $\mathcal{Q}$. The questions are then answered by two LLMs: one that can access all passages $\mathcal{P}$ (RAG) and one that can only access the focus passage p^* (Completion). The two LLMs produce the answers $\mathcal{A}$ and $\mathcal{A}^*$, which are then compared using an embedding model and pairwise cosine similarity.

$$\zeta_{sem} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \bar{\zeta}_{sem}(p, \mathcal{P}) \tag{6}$$

To illustrate semantic independence, consider the following examples demonstrating low and high semantic independence:

Low Semantic Independence

```
#Passage 1
Joe is a professional driver. He is therefore
allowed to drive over the speed limit.

#Passage 2
Exceptions from the speed limits
only apply during the daytime.
```

High Semantic Independence

```
#Passage 1
Joe is a professional driver. He is therefore
allowed to drive over the speed
limit during the daytime.

#Passage 2
Exceptions from the speed limits
only apply during the daytime.
```

In the first example, semantic independence is low because passage 2 contains critical information that modifies the interpretation of passage 1, namely the exception related to daytime. The second example achieves high semantic independence by incorporating all related contexts within passage 1.

3.4 Collective Information Preservation

The preservation of information during document chunking represents a critical requirement as established by principle 3. While previous characteristics (subsection 3.2 and 3.3) evaluate passages from the view of single passages, quantifying information loss demands simultaneous analysis across all passages and the source document. In an ideal scenario, there would exist a comprehensive method $I(\cdot)$ capable of extracting all atomic facts $f \in \mathcal{F}$ from text, where $\mathcal{F}$ encompasses the complete space of factual statements expressible in natural language. This method would extract every atomic fact from the original document D and verify their preservation within the collection of passages $\mathcal{P}$, as formalized in equation 7.

$$\begin{aligned} I : x \to \mathcal{F}_x = \{f_1, f_2, \dots\} \subseteq \mathcal{F} \\ \mathcal{F}_D \subseteq \mathcal{F}_{\mathcal{P}} \end{aligned} \tag{7}$$

However, the implementation of an ideal method $I(\cdot)$ proves impractical for extensive texts due to the inherent complexity of natural language. This complexity makes the complete enumeration of atomic facts virtually impossible. To address this challenge, we propose leveraging a subset of atomic facts $\hat{\mathcal{F}}_D \subseteq \mathcal{F}_D$ as a robust approximation. This subset can be derived through a non-ideal yet practical function $\hat{I}(\cdot)$ that, while not exhaustive, effectively captures key factual information.

$$\begin{aligned} \hat{I} : x \to \hat{\mathcal{F}}_x = \{f_1, f_2, \dots\} \subset \mathcal{F}_x \\ \hat{\mathcal{F}}_D \subseteq \mathcal{F}_D \subseteq \mathcal{F}_{\mathcal{P}} \Rightarrow \hat{\mathcal{F}}_D \subseteq \mathcal{F}_{\mathcal{P}} \end{aligned} \tag{8}$$

As equation 8 delineates, the subset of atomic facts $\hat{\mathcal{F}}_D$ should be fully contained within the space of the chunked sections $\mathcal{F}_C$. The degree of overlap between these sets serves as a decisive indicator of information preservation across the chunking process.

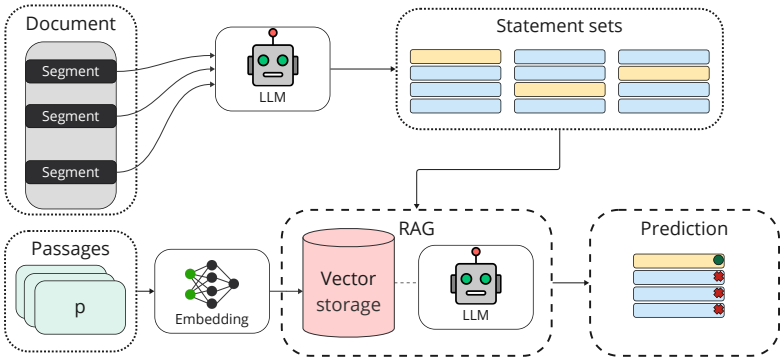


Figure 4: Calculation of the information preservation ζ_{inf}: Three-sentence segments are randomly sampled from the original document. An LLM uses the segments to construct statement quadruplets, where one is verifiable true and three are plausible but false. A secondary LLM is then tasked with truth discrimination by analyzing the statements against contextually relevant passages.

LLMs emerge as an optimal implementation choice for $\hat{I}(\cdot)$, given their exceptional capabilities in semantic comprehension and nuanced fact extraction from natural language [36]. Their demonstrated proficiency in identifying and validating factual relationships within text positions them as ideal candidates for approximating the theoretical information extraction function $I(\cdot)$.

Our method involves sampling document segments d comprising three consecutive sentences from the original document D using a uniform probability distribution $U(\cdot)$. While this localized sampling approach effectively captures information within sentence triplets, we acknowledge its current limitation in detecting complex relationships across longer distances. Each sampled segment undergoes LLM processing to generate quadruples containing one true and three false statements $s_t, s_{f1}, s_{f2}, s_{f3} = q$. The system leverages cosine similarity matching to retrieve relevant passages $\mathcal{P}_{s_t} \subset \mathcal{P}$ from a vector database based on the true statement s_t. These retrieved passages are then analyzed by an LLM tasked with identifying the true statement among the quadruple. To ensure consistent evaluation, the LLM must provide a definitive prediction even under uncertainty. This comprehensive approach is formalized in equation 9 and illustrated in Figure 4.

$$\begin{aligned} d &\sim U(D) \\ LLM(d) &\rightarrow \{s_t, s_{f1}, s_{f2}, s_{f3}\} = s \\ LLM(s, \mathcal{P}_{s_t}) &\rightarrow a \\ \mathcal{A} &= \{a_1, a_2, \dots\} \\ \mathcal{S} &= \{s_1, s_2, \dots\} \\ \zeta_{inf} &= \frac{1}{|\mathcal{S}|} \sum_{a \in \mathcal{A}, s \in \mathcal{S}} \begin{cases} 1 & \text{if } a = s_t \\ 0 & \text{if } a \neq s_t \end{cases} \end{aligned} \tag{9}$$

Our method employs a binary scoring mechanism for ζ_{inf}, prioritizing clarity and interpretability in measuring information preservation. While more nuanced scoring approaches might capture partial information retention, they would introduce unnecessary complexity into the evaluation methodology.

To demonstrate the practical efficacy of our information preservation metric, we present two carefully constructed examples that illustrate the nuanced ways in which chunking strategies can impact information retention. These examples illustrate how subtle variations can produce measurably different outcomes in terms of preserving critical information.

Low Information Preservation

```
#Document
Joe is a professional jogger, therefore he can
finish a marathon in less than 3 hours.

#Passage 1
Joe is a professional jogger.

#Passage 2
Joe can finish a marathon in an impressive time.
```

The first example illustrates a significant degradation in information fidelity, where precision in the information is lost ("less than 3 hours" → "impressive time").

High Information Preservation

```
#Document
Joe is a professional jogger, therefore he can
finish a marathon in less than 3 hours.

#Passage 1
Joe is a professional jogger.

#Passage 2
Joe can finish a marathon in less than 3 hours.
```

In contrast, the second example, the chunking strategy maintains the integrity of all atomic facts present in the original document.

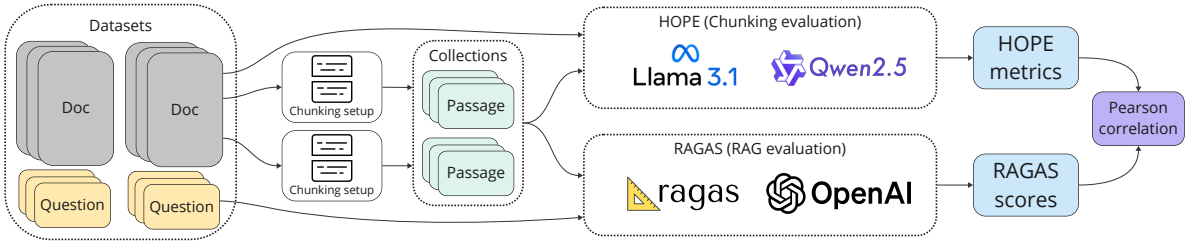


Figure 5: Documents and questions from diverse domains are processed through eight chunking configurations (four fixed-size, two recursive, and two semantic chunking variants). For each resulting passage collection, both the HOPE metric and RAGAS scores are computed. The setup then analyzes correlation between HOPE and RAGAS scores across all configurations and datasets.

This faithful retention of information enables complete reconstruction of the original semantic content, resulting in a superior ζ_{inf} score.

3.5 HOPE Metric

A formal metric rooted in the three principles of chunking (section 1.3), irrespective of domain, and with no need for human annotations, is needed so that more clever chunking methods can be developed and that chunking can be applied more intelligent in real-world RAG systems.

The HOPE metric covers all three levels of passage evaluation (section 3) to provide a holistic assessment of chunking quality. HOPE maps to the interval $[0, 1]$, where 1 represents ideal chunking, and 0 indicates complete chunking failure. The metric is defined as a normalized linear combination of concept unity (subsection 3.2), semantic independence (subsection 3.3), and information preservation (subsection 3.4), as shown in equation 10.

$$HOPE = \frac{1}{3}(\zeta_{inf} + \zeta_{sem} + \zeta_{con}) \tag{10}$$

The implementation of these sub-metrics relies on synthetic data generation from LLMs, which presents specific technical challenges. For robust metric calculation, the generated synthetic questions and answers must satisfy two critical properties: faithfulness and diversity [16]. Insufficient faithfulness results in off-topic or inaccurate questions and statements, potentially based on hallucinated facts, violating the fundamental relationship established in equation 7. Limited diversity in the generated content leads to inadequate representation of the underlying data, constraining the effectiveness of ζ_{con} and ζ_{inf} to a subset of the available information. The broader implications and challenges of utilizing language models for these evaluations are addressed in section 5.

4 Experiment Methodology and Results

We evaluate the effect of HOPE by first collecting a dataset consisting of a diverse set of document types and generating related questions. The documents serve as a knowledge base that we index eight times using distinct chunking setups. For each indexing, both the HOPE metric and the RAG performance indicators are calculated on a corpus level using independent queries. The variance in RAG performance indicators is caused by the effectiveness of the different chunking setups, which is what HOPE intends to model. Thus we will look at correlations between the HOPE metric and the RAG performance metrics to tell how well HOPE can model the effectiveness of chunking.

4.1 Dataset

Existing benchmarks like BEIR [24] and MTEB [18] focus on retriever evaluation, and benchmarks like MMLU [11] target generator evaluation. Passage quality affects both retrieval and generation, thus our dataset needs to address end-to-end RAG system performance. Also, HOPE intends to be domain-agnostic, thus empirical tests must be carried out on documents from different domains and with dissimilar document structures. Following the approach of NQ-open [12], we created seven domain-specific Q&A datasets. For each domain, we collected a set of documents and generated 100 open-book questions using the RAGAS framework. Table 1 presents our diverse document collection, ranging from highly structured technical manuals to more narrative-driven debate transcripts.

Data type	Documents	Source
Newspapers	20	AP, Aljazeera & BBC
Academic articles	12	Arxiv
Wiki articles	54	Wikipedia
Technical manuals	10	VW Car Owners manual
Debates	5	Pile of Law [10]
Terms of service	10	Pile of Law [10]
Medical	20	The COVID-19 Open Research Dataset [26]
Total:	131	

Table 1: An overview of the datasets used for evaluating the HOPE metric.

4.2 RAG Evaluation

RAG performance evaluation is composed of a diverse set of metrics to isolate and test specific aspects of the RAG system. We utilize RAGAS [7], which has developed several LLM-based metrics for evaluating both the final response and the information retrieval components.

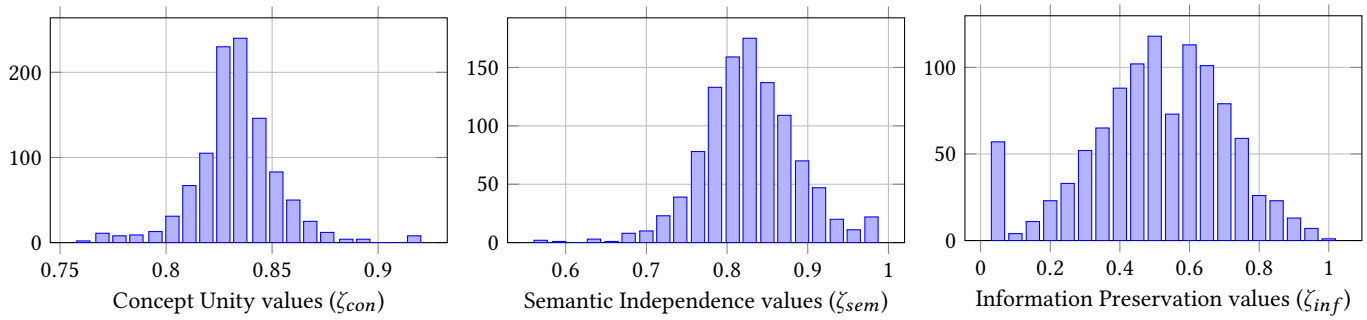


Figure 6: Distributions of the HOPE values for all 1048 combinations of documents and chunking methods.

To explore how the HOPE metrics relate to RAG performance, we selected four metrics that assess both generation quality and retrieval effectiveness. Three metrics analyze information-related characteristics of the response, including semantics, factuality, and structural aspects, while the fourth metric evaluates retrieval quality. The metrics are implemented using the RAGAS python library[1]:

- **Answer Correctness (AC)** Measures answer correctness compared to ground truth as a combination of factuality and semantic similarity.
- **Response Relevancy (RR)** Scores the relevancy of the answer according to the given question. Answers with incomplete, redundant, or unnecessary information are penalized.
- **Factual Correctness (FC)** Uses claim decomposition and natural language inference (NLI) to verify the claims made in the responses against reference texts.
- **Context Recall (CR)** Estimates context recall by estimating TP and FN using annotated answers and retrieved context.

These metrics allow us to evaluate both the effectiveness of information retrieval and the quality of fact preservation across different chunking strategies, directly addressing principles 2 and 3 of chunking.

4.3 Results

Eight different constellations of chunking methods were evaluated to test the versatility of HOPE. These include four fixed-size, two recursive, and two semantic chunking methods. The fixed-size and recursive approaches operate with either large passages (2000 characters) or small passages (500 characters), while the fixed-size method additionally varies the overlap size (500 characters and 125 characters). For semantic chunking, we employed two embedding models: OpenAI's text-embedding-ada-002 [19] and Alibaba's Qwen2 model [14].

We calculated the RAG performance indicators using OpenAI's GPT-4o-mini [21] and text-embedding-ada-002 [19], while the HOPE metrics[2] were computed using NVIDIA's Nemotron-70B [27] (based on Llama 3.1 70B [5]) and Alibaba's Qwen2.5 [14], as illustrated in Figure 5. The distributions of the score for each of the three characteristics across all samples are displayed in Figure 6, with Pearson correlations between RAG and HOPE metrics presented in Table 2. As a baseline comparison, we calculated BLEU scores by comparing the original document with a sequential concatenation of the passages. While BLEU serves as an established metric in NLP evaluations, it provides a conservative baseline for assessing chunking quality.

[1]https://docs.ragas.io/en/stable/

[2]The link to the python implementation of HOPE is removed during the review phase to preserve anonymity, but will be available in the published versions.

	BLEU	HOPE	ζ_{con}	ζ_{sem}	ζ_{inf}
AC	-0.010	0.024	-0.024	**0.105***	0.002
FC	-0.015	0.080*	-0.036	**0.136***	0.053
CR	-0.011	0.011	-0.103*	**0.117***	-0.010
RR	0.027	**0.099***	-0.068*	0.054	0.091*

Table 2: Pearson correlations (ρ) between RAG performance indicators, and HOPE-metrics and BLEU score. All values marked with an asterisk * are statistically significant with p-value < 0.05.

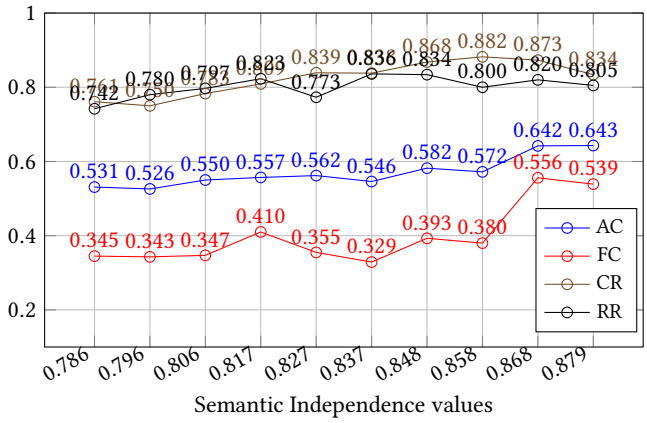


Figure 7: RAG performance indicators plotted against the Semantic Independence values found in Figure 6.

5 Discussion

Our analysis demonstrates that lexical approaches such as BLEU fail to adequately capture chunking quality, as evidenced by their low correlations with RAG performance metrics (Table 2). This limitation is unsurprising given BLEU's fundamental design, which

lacks semantic understanding capabilities [13]. In contrast, HOPE exhibits significant correlations with various RAG performance metrics, particularly through its Semantic Independence component, which shows strong associations with information-based metrics including Answer Correctness and Factual Correctness. The relationship becomes particularly evident in Figure 7, which illustrates consistent improvements across all RAG performance indicators as Semantic Independence values (ζ_{sem}) increase. The magnitude of this effect is substantial, with performance gains of **21.1%** and **56.2%** for Answer Correctness and Factual Correctness, respectively, when comparing readings at minimum and maximum Semantic Independence values. These findings substantiate Semantic Independence as a crucial characteristic in chunking method design, lending empirical support to the second principle of chunking (Principle 2). Moreover, techniques such as decontextualization [8, 20], previously employed in claim decomposition, emerge as promising approaches for enhancing Semantic Independence during chunking.

5.1 Challenging Traditional Assumptions

A particularly intriguing finding emerges from our analysis of Concept Unity values (ζ_{con}), which exhibit negative correlations with all RAG performance indicators (Table 2). This unexpected result challenges the validity or completeness of the first principle of chunking (Principle 1), suggesting that the conventional idea of isolating concepts within passages may be suboptimal. We propose three plausible explanations for this counterintuitive phenomenon: First, the semantic proximity of concepts may prevent the generation of noise in embedding vectors; second, embedding models may be intrinsically optimized for multi-concept passages due to their training data composition; and third, single-concept passages may suffer from reduced information density, potentially diminishing their utility during response generation. The distribution analysis of HOPE metrics (Figure 6) reveals that current Concept Unity implementations generate values within a constrained range ($\approx$ 10% of total range), indicating limited diversity in LLM-generated statements—a recognized limitation in synthetic data generation [16]. While a specialized embedding model for Concept Unity calculation could potentially expand this range, such an approach would contradict HOPE's objective of serving as an automatic metric.

5.2 Limitations and Technical Considerations

The implementation of HOPE as an automatic metric necessitates the integration of LLMs and embedding models, introducing variability and uncertainty, which any language model potentially inherits from biases found in its training corpus [36]. An LLM's bias can lead to less diverse statements in certain areas or topics, thus directly impacting the Concept Unity and Information Preservation components. While LLM performance can be significantly enhanced through sophisticated prompt engineering [30], our current implementation employs relatively straightforward prompts with modest in-context learning examples. More advanced prompt engineering techniques like chain-of-thought reasoning [29] and LLM-based agent networks [33] could potentially improve the statement generation diversity for both Concept Unity and Information Preservation components, but this is left as future work.

5.3 Information Preservation and Response Quality

Our examination reveals that Information Preservation demonstrates a positive correlation with RAG performance (Table 2). The metric exhibits a comprehensive range of values, providing additional validation for the third principle of chunking (Principle 3). Furthermore, the strong correlation between Information Preservation and Response Relevancy suggests a fundamental relationship between information completeness and response quality. This association is theoretically consistent, as Response Relevancy inherently penalizes incomplete information — precisely the phenomenon that Information Preservation aims to quantify.

5.4 Implications and Future Directions

Our findings have substantial implications for the evolution of RAG system architecture. The correlation between Semantic Independence and RAG performance suggests that future system designs should prioritize this characteristic during passage optimization, but more research is needed to verify this empirically. One promising avenue involves leveraging either HOPE or its Semantic Independence component as reward functions within reinforcement learning frameworks for chunk boundary optimization. However, the computational intensity of HOPE, primarily due to multiple LLM invocations, presents significant scalability challenges that warrant further investigation.

Several critical research directions emerge from our analysis. First, the superior performance of multi-concept passages merits deeper investigation, potentially through detailed analysis. Second, the development of computationally efficient chunking strategies that optimize for Semantic Independence while maintaining Information Preservation could yield substantial improvements in RAG system performance. This could deem challenging as Semantic Independence would likely benefit from rewriting the original content, while Information Preservation would benefit from preserving the original structure to avoid "loss in translation".

6 Conclusion

This paper advances the understanding of document chunking in RAG systems through three main contributions. First, we introduce a methodology for characterizing chunking at three separate levels: intrinsic passage properties, extrinsic passage properties, and passages-document coherence. Second, we propose HOPE, a domain-agnostic metric that quantifies these characteristics and provides a systematic approach to evaluating chunking strategies. Third, through empirical evaluation across seven domains, we demonstrate that semantic independence between passages significantly impacts RAG performance, yielding improvements of up to 56.2% in Factual Correctness and 21.1% in Answer Correctness.

Our findings challenge traditional assumptions about Concept Unity within passages, revealing its minimal impact on system performance. These insights provide guidance for optimizing chunking strategies in RAG systems. Future work could explore the adaptation of the HOPE metric to specific domains and investigate its applicability in dynamic chunking scenarios. We believe our methodology lays a solid foundation for systematic improvements of chunking strategies in RAG applications.

References

[1] Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven Failure Points When Engineering a Retrieval Augmented Generation System. section 5, 194–199. ISBN: 9798400705915. DOI: 10.1145/3644815.3644945.

[2] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries Chin-Yew. *Text summarization branches out*, 74–81. DOI: 10.1253/jcj.34.1213.

[3] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 719–729. ISBN: 9798400704314. DOI: 10.1145/3626772.3657834.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Tech. rep. https://github.com/tensorflow/tensor2tensor.

[5] Abhimanyu Dubey et al. 2024. The Llama 3 Herd of Models, 1–92. http://arxiv.org/abs/2407.21783.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, 1–15. http://arxiv.org/abs/2404.16130.

[7] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated Evaluation of Retrieval Augmented Generation. *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 150–158. ISBN: 9798891760912.

[8] Anisha Gunjal and Greg Durrett. 2024. Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification. http://arxiv.org/abs/2406.20079.

[9] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic Models for the First-Stage Retrieval: A Comprehensive Review. *ACM Transactions on Information Systems*, 40, 4, (Oct. 2022). DOI: 10.1145/3486250.

[10] Peter Henderson*, Mark S. Krass*, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of law: learning responsible data filtering from the law and a 256gb open-source legal dataset. (2022). https://arxiv.org/abs/2207.00220.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *ICLR 2021 - 9th International Conference on Learning Representations*.

[12] Kenton Lee, Ming Wei Chang, and Kristina Toutanova. 2020. Latent retrieval for weakly supervised open domain question answering. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 6086–6096. ISBN: 9781950737482. DOI: 10.18653/v1/p19-1612.

[13] Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A Survey on Evaluation Metrics for Machine Translation. *Mathematics*, 11, 4, 1–22. DOI: 10.3390/math11041006.

[14] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning. http://arxiv.org/abs/2308.03281.

[15] Lei Liang et al. 2024. KAG: Boosting LLMs in Professional Domains via Knowledge Augmented Generation, 1–33. http://arxiv.org/abs/2409.13731.

[16] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. ISBN: 9798891760998. http://arxiv.org/abs/2406.15126.

[17] Anurag Mishra. [n. d.] Five Levels of Chunking Strategies in RAG| Notes from Greg's Video — anuragmishra_27746. https://medium.com/@anuragmishra_27746/five-levels-of-chunking-strategies-in-rag-notes-from-gregs-video-7b735895694d. [Accessed 20-11-2024]. ().

[18] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, 2006–2029. ISBN: 9781959429449. DOI: 10.18653/v1/2023.eacl-main.148.

[19] Arvind Neelakantan et al. 2022. Text and Code Embeddings by Contrastive Pre-Training. http://arxiv.org/abs/2201.10005.

[20] Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A Question Answering Framework for Decontextualizing User-facing Snippets from Scientific Documents. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 3194–3212. ISBN: 9798891760608. DOI: 10.18653/v1/2023.emnlp-main.193.

[21] OpenAI. 2023. GPT-4 Technical Report. 4, 1–100. http://arxiv.org/abs/2303.08774.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. DOI: 10.1002/andp.19223712302.

[23] Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*, 11, 36120–36146. DOI: 10.1109/ACCESS.2023.3266377.

[24] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. NeurIPS. http://arxiv.org/abs/2104.08663.

[25] Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. http://arxiv.org/abs/2307.09288.

[26] Lucy Lu Wang et al. 2020. CORD-19: the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics, Online, (July 2020). https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

[27] Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. HelpSteer2-Preference: Complementing Ratings with Preferences, 1–26. http://arxiv.org/abs/2410.01257.

[28] Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A Closer Look at Claim Decomposition, 153–175. ISBN: 9798891761063. DOI: 10.18653/v1/2024.starsem-1.13.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35, NeurIPS, 1–14. ISBN: 9781713871088.

[30] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. http://arxiv.org/abs/2302.11382.

[31] Shangyu Wu et al. 2024. Retrieval-Augmented Generation for Natural Language Processing: A Survey. http://arxiv.org/abs/2407.13193.

[32] Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How Easily do Irrelevant Inputs Skew the Responses of Large Language Models? 1–20. http://arxiv.org/abs/2404.03302.

[33] Zhiheng Xi et al. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. http://arxiv.org/abs/2309.07864.

[34] Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial Report Chunking for Effective Retrieval Augmented Generation. http://arxiv.org/abs/2402.05131.

[35] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. http://arxiv.org/abs/2409.14924.

[36] Wayne Xin Zhao et al. 2023. A Survey of Large Language Models, (Mar. 2023). http://arxiv.org/abs/2303.18223.

[37] Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2024. Mix-of-Granularity: Optimize the Chunking Granularity for Retrieval-Augmented Generation, 1–17. http://arxiv.org/abs/2406.00456.