Investigating the Robustness of Embedding Models on Noisy Input Texts

Michael Dinzinger[1],*, Laura Caspari[1], Kanishka Ghosh Dastidar[1], Jelena Mitrović[1] and Michael Granitzer[1]

[1]*University of Passau, Innstraße 41, 94032 Passau, Germany*

Abstract

Integrating dense retrieval into Retrieval-Augmented Generation systems has significantly enhanced the power of generative language models, especially in the context of conversational search. Despite this advancement, the effect of noisy and poorly structured input text, often drawn from web pages covering multiple topics or languages, has not been thoroughly investigated. This paper addresses that gap by evaluating the robustness of embedding models in handling documents under realistic, noisy conditions beyond the clean and concise golden passage found in most academic datasets. We therefore introduce a synthetic dataset based on the MS-MARCO Passage Collection, in which the original passages are concatenated with text snippets extracted from extraneous web pages. This dataset consists of multiple subcorpora, whose concatenated context varies in length, positioning and relevance to the original query-relevant passage. Our findings reveal two critical limitations to model robustness: first, a positional bias in processing long documents, and second, a significant drop in retrieval performance when models are faced with semantically inconsistent input documents that combine query-relevant information with thematically unrelated (Out-of-Domain) text passages.

Motivated by our observations, we evaluate various chunking methods regarding their ability to preserve contextual and semantic integrity during index creation but find them falling short of the Naive approach in several instances. The paper calls for further research into advanced chunking techniques and alternative retrieval strategies, to improve the robustness of embedding models in RAG frameworks. Our code, datasets and results are made available on GitHub[1] and HuggingFace.[2]

Keywords

Retrieval-Augmented Generation, Dense Retrieval, Text Embeddings, Document Chunking

1. Introduction

Retrieval-Augmented Generation (RAG) has emerged as a powerful framework for enhancing generative language models by incorporating information retrieval techniques [1, 2]. A core component of most RAG pipelines is dense retrieval, which relies on embedding models to retrieve relevant information. However, the performance of these embedding models is typically evaluated on a limited set of benchmarks, such as MTEB [3], whose datasets primarily feature clean and curated documents. These documents contain minimal content beyond the query-relevant information, making them an unrealistic representation of the noisy, heterogeneous data found in real-world applications.

In recent years, web search has been transitioning from traditional result lists to more interactive modes of information retrieval. In such scenarios of conversational search, the web itself acts as a dynamic corpus for retrieving relevant content. This shift introduces new challenges in handling the inherent heterogeneity of web data extracted from various sources. The quality of this corpus is often compromised by suboptimal HTML parsing tools [4]. As conversational search increasingly relies on dense retrievers, the robustness of these retrieval models on low-quality data becomes a significant

[1] https://github.com/michaeldinzinger/chunkeval

[2] https://huggingface.co/michaeldinzinger

Second International Workshop on Open Web Search (WOWS 2025)

*Corresponding author.

michael.dinzinger@uni-passau.de (M. Dinzinger); laura.caspari@uni-passau.de (L. Caspari); kanishka.ghoshdastidar@uni-passau.de (K. G. Dastidar); jelena.mitrovic@uni-passau.de (J. Mitrović); michael.granitzer@uni-passau.de (M. Granitzer)

0009-0003-1747-5643 (M. Dinzinger); 0009-0002-6670-3211 (L. Caspari); 0000-0003-4171-0597 (K. G. Dastidar); 0000-0003-3220-8749 (J. Mitrović); 0000-0003-3566-5507 (M. Granitzer)

concern. Despite their wide application, e.g. in RAG systems, the question of how dense retrievers handle noisy documents – those containing a query-relevant passage alongside extraneous information, either related (In-Domain) or unrelated (Out-of-Domain) – remains underexplored. Therefore, investigating the resilience of dense retrievers to semantically inconsistent web texts is critical for maintaining reasonable performance.

With our work, we aim to fill this gap by studying the impact of noisy input text on the performance of recent embedding models. Through a series of experiments using our tailor-made synthetic dataset, we demonstrate the extent to which text quality and structure affect the accuracy of dense retrieval models. This paper makes the following contributions:

- **Synthetic Dataset:**
  We introduce a tailor-made synthetic dataset based on the MS-MARCO Passage Collection. The dataset includes multiple subcorpora simulating semantically inconsistent web texts.
- **Robustness Analysis of Embedding Models:**
  We present a study of recent embedding models, focusing on their robustness when exposed to noisy input texts that combine query-relevant passages with extraneous information. Our experiments reveal two key constraints:
  - Handling long documents introduces a model-dependent **positional bias**.
  - The presence of **topically unrelated information** (Out-of-Domain context) significantly degrades retrieval performance.
- **Evaluation of Chunking Methods:**
  We compare various chunking methods and assess their ability to preserve contextual coherence when segmenting text documents. However, our findings suggest that existing approaches are either inadequate or impractical, highlighting the need for further research to develop more robust solutions.

The remainder of this paper is organized as follows. In Section 2, we review related work on Retrieval-Augmented Generation and embedding models in dense retrieval. In Section 3, we describe the construction of our synthetic dataset and provide an overview of the embedding models selected for evaluation. Section 4 discusses the experimental results and presents two key limitations of current models. Section 5 compares chunking methods and evaluates their effectiveness in preserving contextual coherence as a potential solution to overcome the models' limitations. Finally, Section 6 concludes the paper with a summary of our contributions and a discussion of open research questions.

2. Related work

2.1. Dense retrieval in RAG systems

Dense retrieval originated from advancements in deep learning, particularly in the application of neural networks to map queries and documents into dense vector representations using SentenceTransformers [5]. In many cases, such models allow for more effective and scalable information retrieval compared to traditional sparse methods like term-based matching. Well known dense retrieval techniques are, e.g., Contriever [6] and DPR [7]. These two as well as more recent embedding models applicable in RAG systems, such as the series of Snowflake Arctic models [8], are pre-trained, fine-tuned and/or benchmarked on the prominent BEIR collection [9]. The respective datasets contain mostly clean and concise passages, which are on average shorter than 512 tokens and thus small enough to fit in the context window of most common embedding models. Hence, most prior work assumes clean and well-structured documents, whereas real-world scenarios involve noisy, heterogeneous web data. The impact of such noise on dense retrievers remains underexplored, and our work addresses this gap by systematically generating controlled noisy data.

2.2. The *Power of Noise*

Investigating further on the cleanliness of text chunks, Cuconasu et al. study the *Power of Noise* [10]. Their work analyzes which types of documents a RAG system should retrieve for effective prompt formulation, regarding the relevance of the retrieved documents to the query, their position and the number included in the context. The authors cannot conclude to any clear strategy for integrating retrieved documents into the input prompt of generative language models. However, they observe that accuracy is higher when the gold document is positioned near the query, lower when it is placed furthest from the query, and lowest when it appears in the middle of the context. Interestingly, even the inclusion of irrelevant chunks can unexpectedly boost accuracy by more than 30%.

2.3. Study of retrieval units

Previous works on dense retrieval consider paragraphs as a reasonable retrieval unit. Recent studies in the context of RAG, however, explore very short as well as long retrieval units, both as seemingly effective measures to increase the performance of RAG systems in regards to their Question-Answering capabilities. On the one hand, Chen et al [11] introduce *propositions* as the most fine-granular unit of segmenting and indexing the retrieval corpus. Propositions are defined as atomic and self-contained expressions, each encapsulating a distinct factoid in natural language. The authors conclude that selecting the proper retrieval granularity for a given query at inference time, including propositions, sentences and passages, can improve the dense retrievers' performance.

On the other hand, Jiang et al [12] propose a new framework called *LongRAG*, consisting of a long-context retriever and a long-context reader. The authors want RAG systems to fully leverage the long-context capabilities of recent LLMs. First, by extending the retrieval granularity up to documents or even groups of documents, the number of elements in the corpus decreases drastically, helping the retrieval performance. Second, as the entire original document, e.g. the main content of a web page, is put in the context of the LLM, the model performs well on multi-hop questions due to its extensive capabilities to reason over long input text. Instead of leaving the task of finding the "needle in the haystack" to the retriever, the authors pledge for long-context LLMs to do the job and extract the relevant information from the retrieved text in order to answer the user query accurately.

Another recent approach that contributes to the study of retrieval units is *Landmark Embeddings*, introduced by Luo et al. [13]. Unlike conventional chunking, which breaks the text into discrete segments and potentially loses contextual coherence, Landmark Embeddings use a chunking-free strategy by attaching a special token at the end of each sentence, called a landmark. These tokens capture the semantic information of the sentence and its surrounding context, preserving a more unified representation of long texts. This method again leverages the long-context capabilities of recent LLMs, to improve the quality of dense retrievals without sacrificing context. However, both approaches, LongRAG and Landmarks Embeddings, operate under the assumption of relatively clean input data and do not systematically evaluate robustness against heterogeneous or noisy documents as encountered in real-world conversational search.

2.4. The *Dwell in the Beginning* effect

Coelho et al. study the existence of a *Dwell in the Beginning* effect in Transformer-based models [14]. Dwell in the Beginning describes a positional bias in text representation learning, particularly in the context of web document retrieval, in which vector embeddings of long documents tend to capture information located at the beginning of the input text better. The authors observe that this phenomenon starts emerging during unsupervised contrastive pretraining, and furthermore heavily relies on the data distribution during fine-tuning. Coelho's findings emphasize that embedding quality for long inputs depends on various – some even counter-intuitive – factors. One such factor, among others, is the positioning of the relevant information within the input sequence. This positional bias further motivates our investigation into how the placement of relevant information in noisy contexts affects retrieval performance.

2.5. Chunking strategies

During indexing, the first phase of a vector-based RAG pipeline, documents are processed, segmented, and transformed into embeddings to be stored in a vector database. In this context, chunking refers to the process of breaking down input text into smaller, manageable pieces [15]. This is often necessary to ensure that long documents are split into portions that fit in the embedding model's token limit. However, chunking does more than just accommodate model constraints; it also enhances contextual coherence within each segment.

Naive chunking methods, such as splitting the text into fixed-length segments, are widely used because of their simplicity and effectiveness in many cases. These methods can be enhanced by introducing chunk overlap or by setting breakpoints only at sentence boundaries. Furthermore, Günther et al presented an innovative approach called Late Chunking, where text is split after being processed by the Transformer model but before mean pooling [16]. This technique takes advantage of Transformer models' long-context capabilities, and ensures that each chunk captures the broader document's context within its embedding.

Beyond naive approaches, more advanced techniques like Semantic Chunking, Agentic Chunking [17], and Contextual Retrieval [18] have emerged using LMs in the chunking process. Such methods may include dynamic breakpoints, where text is divided based on semantic coherence rather than fixed length, or augmenting chunks with summarized information from neighboring segments. The goal of these approaches is to preserve both the contextual richness and semantic integrity of the text, resulting in embeddings that are better aligned with the retrieval task. Yet, while these methods improve contextual coherence, it remains unclear to which degree they are capable of handling semantically inconsistent and noisy web documents, which we systematically simulate in our synthetic dataset.

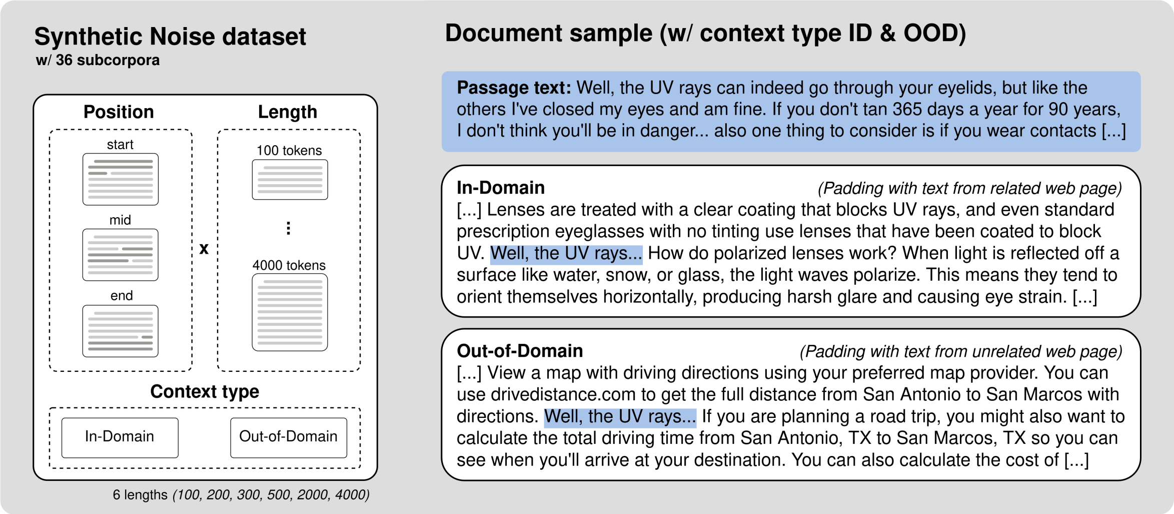
3. Experimental Setup

3.1. Synthetic dataset

To evaluate the robustness of embedding models and chunking methods under real-world conditions, we created a synthetic dataset designed to simulate "noisy" documents. This noise is modeled to intentionally break contextual coherence, in order to study its impact on retrieval performance. For instance, we observe that, after applying respective parsing tools, text snippets extracted from web documents still contain disconnected content such as user comments or boilerplate sections. Web texts hence remain inconsistent in their structure and content, combining text segments which are not necessarily related to each other.

Figure 1 illustrates the composition of the Synthetic Noise dataset, which consists of 41,371 document samples per subcorpus. It is constructed using passages from the MS-MARCO collection [19], augmented by adding artificial context. The original passages are concise, averaging about 72 tokens (measured using the BERT tokenizer), and typically, these snippets contain only the information directly relevant to their corresponding query. In addition to the passage text, the MS-MARCO dataset includes URLs linking to the source web pages, allowing us to scrape further content related to the original piece of information. The Synthetic Noise samples are constructed by padding the given short passages with extraneous content from web pages linked in the MS-MARCO collection. These synthetic documents vary along three key dimensions, resulting in 36 distinct subcorpora derived from the 41,371 original core passages:

- **Document length:**
  We vary the length of the synthetic documents to measure how the ratio of relevant to query-irrelevant information impacts retrieval performance.
- **Positioning of the original passage:**
  We adjust the position of the original text snippet within the document to study positional biases in its dense vector representation.


Synthetic Noise dataset
w/ 36 subcorpora
Position
start
mid
end
x
Length
100 tokens
⋮
4000 tokens
Context type
In-Domain
Out-of-Domain
6 lengths (100, 200, 300, 500, 2000, 4000)
Document sample (w/ context type ID & OOD)
In-Domain
(Padding with text from related web page)
Out-of-Domain
(Padding with text from unrelated web page)

Passage text: Well, the UV rays can indeed go through your eyelids, but like the others I've closed my eyes and am fine. If you don't tan 365 days a year for 90 years, I don't think you'll be in danger... also one thing to consider is if you wear contacts [...]

[...] Lenses are treated with a clear coating that blocks UV rays, and even standard prescription eyeglasses with no tinting use lenses that have been coated to block UV. Well, the UV rays... How do polarized lenses work? When light is reflected off a surface like water, snow, or glass, the light waves polarize. This means they tend to orient themselves horizontally, producing harsh glare and causing eye strain. [...]

[...] View a map with driving directions using your preferred map provider. You can use drivedistance.com to get the full distance from San Antonio to San Marcos with directions. Well, the UV rays... If you are planning a road trip, you might also want to calculate the total driving time from San Antonio, TX to San Marcos, TX so you can see when you'll arrive at your destination. You can also calculate the cost of [...]

Figure 1: Synthetic document samples with In-Domain and Out-of-Domain context.

- **Context type:**
  We include both related and unrelated contexts to assess how semantic consistency in the input text (or lack thereof) influences retrieval performance.

Our dataset thus includes samples with concatenated contexts in two distinct levels of relevance to the core passage. In-Domain contexts are extracted from web pages that contain information related, though not necessarily relevant to the query associated with the original passage[1]. While these contexts do not directly address the query, they often discuss similar topics. This maintains – to a certain extent – contextual coherence within the synthetic document. In contrast, Out-of-Domain contexts are taken from web pages randomly selected from the MS-MARCO corpus and lack any objective relevance to the original query or passage. Since these snippets typically cover entirely unrelated subjects, they break the contextual coherence. The lack of coherence makes it more difficult for embedding models to accurately represent the relevant information of the original passage in the document vector.

As discussed, web texts often lack consistent contextual coherence. Using our synthetic dataset, we systematically examine the impact of document length, positional bias, and the consistency of context relevance on embedding models. By isolating these factors, we aim to understand how coherence deteriorates and how noisy, real-world data affects both retrieval accuracy and efficiency.

Table 1
List of evaluated embedding models

Embedding Model	Model Size (M Parameters)	Dimensionality	Max Tokens	NDCG@10 on MS-MARCO
BGE small (en, v1.5) [20]	33	384	512	40.83
Arctic M (v1.5) [8]	109	768	512	42.03
Stella (1.5B, en, v5)	1,543	1,024	512	45.22
Jina (v3) [21]	572	1,024	8,192	40.82
GTE Qwen2 (1.5B) [22]	1,776	1,536	32,000	43.36
GTE Qwen2 (7B) [22]	7,613	3,584	32,000	45.98

[1] URLs are considered as In-Domain if marked as “not selected” in the MS-MARCO collection. Selecting the original, selected web pages however yields similar results.

Table 2
nDCG@10 in % on the Synthetic Noise dataset. Its subcorpora vary along 3 dimensions: context type, length, and position. The ratio of relevant information to ID/OOD context is given as percentage after the document length. Bold values indicate the highest performance within each subtable (i.e., for a specific context type and length group), while underlined values mark the highest score within each row (i.e., for a specific model and context type across different positions).

Embedding model	Context type	100 tokens (68%)			300 tokens (24%)			500 tokens (14%)		
		start	mid	end	start	mid	end	start	mid	end
BGE small	ID	**78.54**	77.39	77.00	**74.88**	66.49	61.76	**75.52**	62.51	60.68
baseline: 77.69	OOD	70.69	61.54	59.71	33.87	11.55	7.89	25.02	2.28	4.41
Arctic M	ID	**79.97**	79.32	79.85	**81.48**	72.91	70.50	**80.01**	68.17	62.45
baseline: 79.64	OOD	73.00	65.74	66.04	54.15	13.05	18.57	46.45	6.40	4.28
Stella	ID	**81.90**	79.72	78.44	**80.30**	75.11	73.11	**78.82**	74.13	67.11
baseline: 81.95	OOD	72.80	56.94	53.26	42.69	10.82	17.48	30.48	7.68	8.01
		100 tokens (68%)			500 tokens (14%)			4000 tokens (2%)		
Jina	ID	**77.03**	74.59	74.81	**75.65**	67.64	70.10	59.69	**61.94**	54.14
baseline: 77.33	OOD	68.24	56.79	58.94	31.23	6.41	19.68	2.98	1.64	1.67
GTE Q. 1.5B	ID	78.31	**78.49**	78.21	74.80	74.48	**76.47**	68.83	69.91	**77.64**
baseline: 77.84	OOD	63.86	60.01	59.56	11.32	6.02	25.93	6.17	1.84	33.20
GTE Q. 7B	ID	**83.83**	82.44	80.90	**81.64**	77.38	76.98	74.80	73.68	**78.60**
baseline: 83.04	OOD	77.44	66.48	59.90	37.44	13.34	23.01	17.83	6.56	33.85

3.2. Embedding models

Table 1 lists six dense retrieval models in ascending order based on their performance, dimensionality, and number of parameters. The selected embedding models fall into two groups: 512-token models and long-context models (8,192+ tokens), both widely used as first-stage retrievers in RAG systems due to their strong performance on academic benchmarks like the Massive Text Embedding Benchmark (MTEB) [3]. The 512-token models are efficient for shorter, more structured documents. In contrast, long-context models are designed to process extensive input sequences, making them better suited for handling large, fragmented web documents. Recent works such as Late Chunking [16] and Long-RAG [12] demonstrate the potential that long-context retrievers offer to RAG, despite the costs of greater model sizes and computational overhead.

4. Robustness of Embedding Models

This section studies the robustness of embedding models as retrievers when exposed to documents containing extraneous information varying in length, positioning and relevance to the original answer text. Our experiment tested the performance of six models (listed in Table 1) on the Synthetic Noise dataset (described in Section 3.1). Each subcorpus in the dataset consists of 41,371 synthetic documents as well as additional passages from the MS-MARCO collection, increasing the total corpus size to 200,000 documents. These supplementary documents are not relevant to any of the 41,371 queries, hence they introduce further difficulty to the retrieval task without contributing relevant samples.

To effectively measure the impact of noise, the models were evaluated on a downstream retrieval task, closely following evaluation setups like those used in MTEB [3]. The primary performance metric used was nDCG@10.

Retrieval results: In Table 2, we observe that when evaluating on datasets containing synthetic documents of 100 tokens, the performance does not significantly deviate from the baseline. The baseline retrieval performance is provided by the core passages without any dispensable information. Interestingly, for four out of six models, performance with 100-token documents slightly improves over

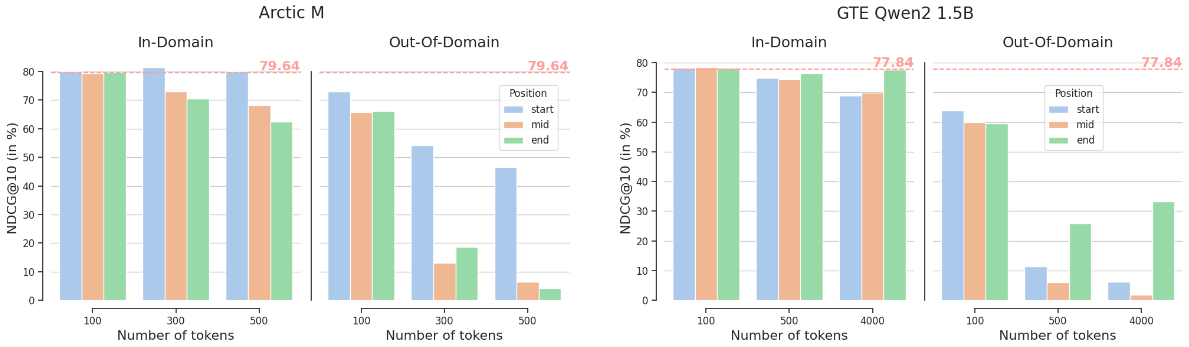


Figure 2: Visual comparison of retrieval results measured as NDCG@10. The red baseline refers to the corpus of original clean passages.

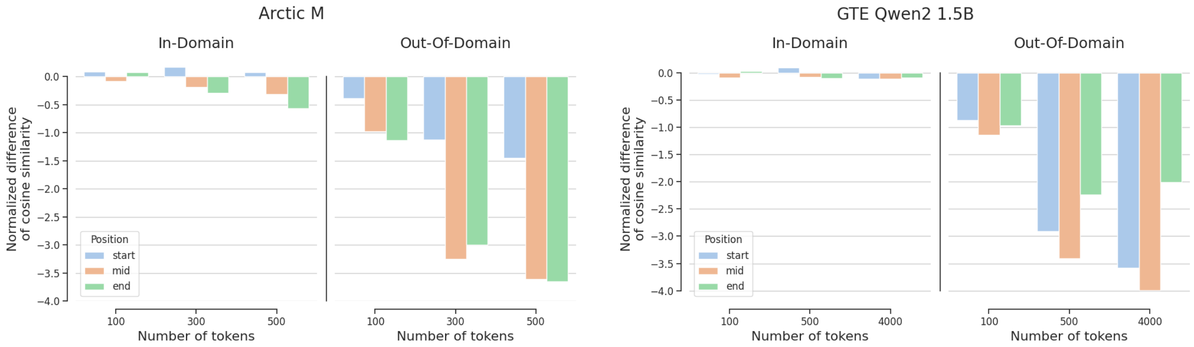


Figure 3: Normalized differences of similarity values after noisy context is added.

the baseline. This observation regarding embedding models aligns with the findings of Cuconasu et al [10], who concluded that the influence of noise in the input context of generative language models is not yet fully understood and may, under certain conditions, even improve performance.

As the length of the synthetic documents increases, the ratio of relevant information to manually added context becomes smaller. For instance, in a 4,000-token document, the original relevant passage accounts for only around 2% of the total text, meaning that around 98% of the input consists of potentially irrelevant information. Such an overwhelming presence of noise can significantly distract the embedding models from the query-relevant snippet, which is expected to be well represented in the final dense vector. Consequently, this vast amount of noise inevitably leads to a drop in performance on downstream retrieval tasks.

Consequently, for most of the models nDCG@10 declines as documents become longer. While the performance drop is relatively moderate when the noise is thematically related (In-Domain) to the core passage, it becomes much more significant when the context is unrelated. OOD noise tends to disrupt the semantic integrity of the document, making it harder for the models to retain and represent the key information accurately. This suggests that the relevance of the context is a critical factor in determining the models' ability to preserve the semantics of the query-relevant information amidst noise.

Visualizations: Furthermore, we conduct an in-depth comparison between the Arctic M and GTE Qwen2 1.5B models, illustrated in Figures 2 and 3. Arctic M, with a 512-token limit, and GTE Qwen2 1.5B, optimized for long-context processing, exemplify their respective groups in Table 1 and demonstrate good overall performance. Figure 2 depicts the nDCG@10 scores for both models, visualizing the numeric retrieval metrics of Table 2. Beyond that, Figure 3 visualizes the normalized differences in similarity values, defined as:

$$\text{nDiff}(doc) = \frac{s_{p_{doc}} - s_{doc}}{\text{std}(S_{p_{doc}})}$$

where s_{doc} represents the cosine similarity between the query and the noisy document embedding. Similarly, $s_{p_{doc}}$ is the cosine similarity[2] between the query and the original passage embedding, and $\text{std}(S_{p_{doc}})$ is the standard deviation of similarity values across the corpus. By dividing by $\text{std}(S_{p_{doc}})$, we apply a paired Z-score normalization. Since different models may interpret similarity with varying internal scales, this normalization ensures that differences are comparable across models, regardless of the specific range of their cosine similarity scores. Hence, it makes the comparison across models more consistent.

The analysis of Table 2 together with Figures 2 and 3 uncover two noteworthy observations:

Positional bias: A significant positional bias is observed, though the nature of this bias varies between models. For Arctic M, the normalized similarity differences in Figure 3 are consistently higher when the query-relevant information is positioned at the beginning of the document. This model relies on absolute positional encoding, a standard approach for Transformer models. In contrast, GTE Qwen2 1.5B [22] employs a more advanced relative positional encoding system, which allows it to better handle longer sequences. Interestingly, for this model the smallest similarity difference is found when the query-relevant information is positioned at the end of the document. GTE Qwen2 1.5B thus reverses the "Dwell in the Beginning" effect [14], which refers to the bias towards the beginning of a document in traditional Transformers. In both models, the magnitude of the positional bias increases as document length grows. This observation is consistent with the higher nDCG@10 scores at the respective positions in Figure 2, to which the positional bias applies.

Context relevance: In Figure 3, the differences in similarity remain relatively small for In-Domain noise, but become an order of magnitude larger for Out-of-Domain noise. This is again consistent with the observed drop in retrieval performance for OOD contexts, depicted in Figure 2. Notably, in the presence of In-Domain noise, both models perform near the baseline, even for long documents approaching their maximum context length. This highlights the models' ability in managing coherent inputs of dynamic lengths. But it also demonstrates the major constraint of this robustness, namely their inability in handling input texts combining query-relevant and unrelated information.

Even though this observation is entirely intuitive, the extent to which retrieval performance declines, is nevertheless remarkable. For example, nDCG@10 scores for Arctic M drop by at least 33 points for 500-token documents, and at least 50 points for GTE Qwen2 1.5B for documents of the same size.

5. Document Chunking

In the previous section, we identified two critical limitations in the robustness of embedding models: positional biases and the models' inability in keeping good performance despite the existence of query-irrelevant information in the input documents. In the last part of our paper, we turn our attention to techniques that can potentially overcome these constraints. Such techniques may enhance retrieval performance in RAG systems by creating contextually more coherent text documents during the indexing phase. One of the most prominent approaches for optimizing index construction is the use of effective chunking methods [23].

In the following study, we evaluate how well current chunking strategies mitigate the impact of noisy input texts. We compare full document retrieval (i.e., no chunking) against three popular chunking techniques on several benchmark datasets using the Jina (v3) embedding model. The first four datasets listed in Table 3 are standard datasets used in the MTEB Retrieval Evaluation. In addition, the study includes the MS-MARCO Document Collection[3] [19] and the NarrativeQA dataset from the LongEmbed benchmark [24], both of which contain longer documents than typical academic retrieval benchmarks. For the final dataset, we sampled 41,371 documents from the 18 OOD subcorpora of our Synthetic Noise dataset.

[2] We repeated the analysis using dot product, yielding nearly identical results.

[3] To reduce comput. costs, we use a randomly sampled subset of 300,000 documents.

Table 3
Impact of chunking methods on nDCG@10 in %. #Tokens is the average number of tokens per document using a BERT-based Tokenizer. Embedding Model: Jina (v3)

Dataset	#Tokens	Full	Naive	Late	Semantic	Optimal
FiQA2018	176.2	47.33	46.21	**47.51**	42.60	-
NFCorpus	337.1	36.69	35.54	**36.80**	35.18	-
SciFact	316.5	72.33	71.83	**73.21**	71.09	-
TREC-COVID	318.4	**77.68**	72.95	77.52	70.85	-
MS-MARCO Doc. (300k)	1,604.5	70.18	**73.01**	71.61	-	-
NarrativeQA (LongEmbed)	74,843.6	34.25	**73.99**	40.56	-	-
Synth. Noise samples (OOD)	2,106.2	22.93	44.45	41.63	34.35	**78.61**

Fixed-length text splitting: In Table 3, we observe that choosing the appropriate chunking strategy can significantly impact retrieval performance. For short-document datasets, naive chunking – splitting the text based on a fixed token length – performs worse than full document retrieval. This is because naive chunking often results in the loss of important contextual information across chunks, thereby reducing retrieval accuracy. Late Chunking provides a solution in these cases by preserving contextual information from surrounding chunks, leading to noticeably higher retrieval scores.

However, for datasets with longer documents, such as MS-MARCO Document, NarrativeQA, and Synthetic Noise, the trend reverses. These datasets contain documents that typically span beyond one or two paragraphs, and under such conditions, naive chunking outperforms full document retrieval as well as more advanced methods. Interestingly, it even slightly surpasses the performance of Late Chunking. For longer documents surpassing 1000+ tokens, it is seemingly neither effective to embed the full document nor to apply chunked mean pooling, as the contextual information captured from the entire document dilutes the semantic vector representation.

Semantic Chunking: Our study furthermore demonstrates that there is currently no chunking strategy that serves as a one-size-fits-all solution, and while advanced methods may offer theoretical advantages, they are not always practical in real-world applications. For example, Semantic Chunking, which involves dynamically setting breakpoints based on semantic similarity thresholds, is computationally expensive and highly sensitive to hyperparameter tuning. We used, for instance, the NLTK Sentence Splitter, applied Jina (v2, small) to score sentence similarity and tried three different breakpoint thresholds[4]. But in our case – as well as in other scenarions where fine-tuning is difficult or impractical –, Semantic Chunking often fails to come up with reasonable breakpoints between text segments, and hence also fails to achieve the expected improvements in contextual coherence. Furthermore, due to its high computational costs, we could not evaluate Semantic Chunking on the extensive long-document datasets.

Optimal text splitting: When analyzing the Synthetic Noise samples, we can precisely identify the query-relevant passages that were injected into the synthetic documents, allowing us to benchmark retrieval performance against ideal text splitting. In such a scenario, when text chunking is perfectly applied to isolate query-relevant information, it significantly enhances retrieval performance. When compared to existing chunking methods, the optimal approach achieves a substantial improvement, with over 30% higher nDCG@10 scores.

[4] We experimented with the 50^{th}, 70^{th} and 95^{th} percentile as breakpoint thresholds for similarity values, with the median giving the best results (listed in Table 3).

6. Conclusion

In this paper, we have explored the robustness of embedding models within the framework of RAG, focusing particularly on the challenges posed by noisy, heterogeneous web data. In line with Coelho et al. [14], our results confirm the presence of positional biases on longer input texts. However, we show that these biases have a stronger negative impact when combined with query-irrelevant content, a scenario typical of web data, but less explored in prior work. This negative impact leads to a drop in retrieval performance of at least 33 points in nDCG@10 across all embedding models.

We further investigated document chunking as a potential solution to these issues by evaluating several chunking strategies across multiple benchmarks, including long-document datasets and subcorpora samples from our Synthetic Noise dataset. Our experiments revealed that no single strategy is universally effective. Existing approaches either fail to produce semantically rich chunks or remain impractical due to the need for extensive hyper-parameter tuning and high computational costs.

Surprisingly, while prior works proposed advanced chunking strategies (e.g., Late and Semantic Chunking) as solutions, our experiments reveal that these methods often fail to outperform simpler Naive Chunking when applied to noisy data. Nevertheless, these recent ideas lay the conceptual foundation for future exploration towards more advanced chunking and retrieval techniques. For instance, the recent concept of Landmark Embeddings introduces a chunking-free approach that leverages large language models' long-context capabilities. Combining these approaches with innovative techniques, such as Contextual Retrieval [18], may provide more robust solutions to the challenges posed by semantically inconsistent web texts in retrieval-augmented systems.

Acknowledgments

This work is part of OpenWebSearch.eu, funded by the EU under GA 101070014, and part of CAROLL, funded by the German Federal Ministry of Education and Research (BMBF) under 01|S20049.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Improve writing style, Paraphrase and reword. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[2] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, Z. Liu, Evaluation of Retrieval-Augmented Generation: A Survey, 2024. doi:10.48550/ARXIV.2405.07437.

[3] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive Text Embedding Benchmark, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2023. doi:10.18653/v1/2023.eacl-main.148.

[4] J. Bevendorff, S. Gupta, J. Kiesel, B. Stein, An Empirical Comparison of Web Content Extraction Algorithms, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, ACM, 2023. doi:10.1145/3539618.3591920.

[5] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, 2019. doi:10.48550/ARXIV.1908.10084.

[6] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised Dense Information Retrieval with Contrastive Learning, 2021. doi:10.48550/ARXIV.2112.09118.

[7] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.550.

[8] L. Merrick, D. Xu, G. Nuti, D. Campos, Arctic-Embed: Scalable, Efficient, and Accurate Text Embedding Models, 2024. doi:10.48550/ARXIV.2405.05374.

[9] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, 2021. doi:10.48550/ARXIV.2104.08663.

[10] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, F. Silvestri, The Power of Noise: Redefining retrieval for rag systems, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, ACM, 2024. doi:10.1145/3626772.3657834.

[11] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, D. Yu, Dense X Retrieval: What Retrieval Granularity Should We Use?, 2023. doi:10.48550/ARXIV.2312.06648.

[12] Z. Jiang, X. Ma, W. Chen, LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs, 2024. doi:10.48550/ARXIV.2406.15319.

[13] K. Luo, Z. Liu, S. Xiao, K. Liu, BGE Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models, 2024. doi:10.48550/ARXIV.2402.11573.

[14] J. Coelho, B. Martins, J. Magalhães, J. Callan, C. Xiong, Dwell in the Beginning: How language models embed long documents for dense retrieval, 2024. doi:10.48550/ARXIV.2404.04163.

[15] B. Smith, A. Troynikov, Evaluating Chunking Strategies for Retrieval, Technical Report, Chroma, 2024. URL: https://research.trychroma.com/evaluating-chunking.

[16] M. Günther, I. Mohr, B. Wang, H. Xiao, Late Chunking: Contextual Chunk Embeddings Using Long-Context Embedding Models, 2024. doi:10.48550/ARXIV.2409.04701.

[17] G. Kamradt, 5 Levels of Text Splitting, https://github.com/FullStackRetrieval-com/RetrievalTutorials, 2024. Accessed: 2024-01-10.

[18] Anthropic AI, Introducing contextual retrieval, https://www.anthropic.com/news/\contextual-retrieval, 2024. Accessed: 2024-01-10.

[19] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, T. Wang, MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2016. doi:10.48550/ARXIV.1611.09268.

[20] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-Pack: Packed Resources For General Chinese Embeddings, 2023. doi:10.48550/ARXIV.2309.07597.

[21] S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, N. Wang, H. Xiao, jina-embeddings-v3: Multilingual Embeddings With Task LoRA, 2024. doi:10.48550/ARXIV.2409.10173.

[22] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, M. Zhang, Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023. doi:10.48550/ARXIV.2308.03281.

[23] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented Generation for Large Language Models: A Survey, 2023. doi:10.48550/ARXIV.2312.10997.

[24] D. Zhu, L. Wang, N. Yang, Y. Song, W. Wu, F. Wei, S. Li, LongEmbed: Extending Embedding Models for Long Context Retrieval, 2024. doi:10.48550/ARXIV.2404.12096.