AMSI SUMMERRESEARCH SCHOLARSHIPS 2025–26

Get a taste for Research this Summer

Investigating the Topic Capacity of Document Embeddings

Tristan Trieu

Supervised by Dr Laurence Park

Western Sydney University

Jane Street®

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

Abstract

The primary aim of this project is to investigate how multiple topics are represented within document embeddings generated by large language models (LLMs), with particular focus on identifying the point at which embedding quality begins to degrade as topic complexity increases. Since direct numerical interpretation of embedding vectors does not provide meaningful insight into topic representation, document retrieval accuracy is adopted as a quantitative evaluation for how consistently topic information is preserved as the number of topics per document increases. Retrieval performance is assessed using established information retrieval metrics. It is hypothesized that retrieval accuracy decreases as topic multiplicity increases, reflecting inherent limitations in single-vector embedding representations. The findings contribute to a deeper understanding of the representational capacity of document embeddings in retrieval and clustering contexts, and provide a foundation for future research into large language models and their applications.

1 Introduction

Recent advancements in large language models have led to the widespread use of document embeddings in information retrieval and related tasks. These embeddings represent textual data as fixed-length vectors in high-dimensional space, enabling efficient semantic comparison and ranking. However, as retrieval systems increasingly depend on dense representations, questions arise regarding their representational limits. This project investigates the topic capacity of document embeddings by analyzing how retrieval effectiveness varies as the number of topics within a document increases. Using established information retrieval metrics, the study seeks to determine whether there exists a measurable point at which embedding quality degrades, thereby clarifying the limitations of single-vector representations and informing future topic-aware retrieval approaches.

1.1 Background

The advent of the Industry 4.0 era, characterized by rapid digitalization and exponential growth in online information has fundamentally transformed how information is generated, stored, disseminated, and accessed. Search engines were then becoming a necessity in providing structured information retrieval and organization, giving rise to the development of contemporary web search systems (Veluru et al. 2025).

1.2 Web Search

Web search refers to the process of retrieving information from a large collection of documents in response to a user's query. Early web search systems relied solely on keyword matching techniques, in which documents are retrieved based on lexical overlap with query terms (Veluru et al. 2025). This approach proved effective when web-scale data was fairly limited back in the days; however, it struggled with contextual diversity and synonymy, preventing it from fully represent semantic connections between queries and documents (Aghaei et al. 2022).

To address these shortcomings, research in information retrieval has shifted towards semantic modelling and machine learning-based ranking frameworks (Veluru et al. 2025). Recent advancements include deep neural architectures, particularly transformer-based models to generate dense vector representations of text (Aghaei et al. 2022). These vector representations are also commonly referred to as embeddings, which encode contextual and semantic information in continuous space, enabling similarity comparison beyond surface-level term overlaps (Veluru et al. 2025). This new approach has improved Web Search accuracy by involving semantic relevance prediction through modelling conceptual similarity between queries and documents (Aghaei et al. 2022).

1.3 Document Embeddings

Document Embeddings are generated by inputting next into a large language model, in which textual data is converted into fixed-length continuous vectors where each dimension represents latent semnatic features learned from large corpora. Once generated, embeddings are organized in a high-dimensional vector space that semantically similar vectors are located closer together according to a specific distance metric, commonly cosine similarity. This spatial alignment reflects the degree to which different documents might refer to the same topics or not. With this form of clustering in the vector space, many applications are made possible including information retrieval and unsupervised grouping based on geometric proximity. According to Gómez & Vázquez (2022), the effectiveness of similarity measures and clustering algorithms depends directly on the quality of document embeddings and their geometric alignment within vector space.

Statement of Authorship

I declare that this report is my own work and has not been submitted for any other academic award. All sources of information, including published and unpublished works, have been acknowledged appropriately. This research was conducted under the supervision of Dr Laurence Park as part of the AMSI Summer Research Scholarship 2025–26. While general guidance and feedback were provided, all experimental design, implementation, analysis, and writing were carried out by the author. Limited code execution and data processing were conducted with technical assistance from the supervisor due to hardware constraints. Generative AI tools were used for language refinement, coding assistance, and knowledge exploration. All intellectual contributions, research decisions, and conclusions remain the author's original work.

2 Research Question

2.1 Literature Review

Recently, it has been shown that representing an entire document as a single fixed-dimensional vector may have inherent limitations. Kesiraju et al. (2020) has addressed this loss by claiming that a document embedding should not be treated as a single point estimate, but rather, as Gaussian distributions (more commonly known as Normal distribution) where the covariance implies uncertainty about the embedding and can be exploited for downstream topic identification. In 2020, Khattab and Zaharia propose ColBERT, which mitigates single-vector compression by representing queries and documents as contextualized token embeddings and computing relevance via late aggregation, enabling fine-grained matching while retaining efficiency through offline document encoding (Khattab & Zaharia 2020). Another innovation was also introduced recently by Chatterjee and Dalton during SIGIR 2025[1], where they proposed QDER, a query-specific multi-vector re-ranking framework that preserves token and entity-level representations throughout the scoring process before performing late interaction.

2.2 The Problem

Despite prior efforts recognising that embeddings possess finite representational capacity, and alternative approaches aimed at mitigating semantic loss through multi-vector architectures, the fundamental capacity limit of single-vector document embeddings remains unquantified. In particular, the extent to which embedding quality degrades as semantic complexity increases has not been systematically examined. This motivates the central research questions of this study:

- What is the representational capacity of a single document embedding?
- How does increasing the number of topics within a document affect web search retrieval effectiveness?

As there is no direct metric for measuring the amount of semantic information encoded within an embedding vector, retrieval effectiveness is adopted as an indirect quantitative proxy. Specifically, changes in web search accuracy are analyzed as the number of topics within each document increases. It is hypothesized that retrieval performance will decline as semantic complexity compounds, reflecting progressive information compression within fixed-dimensional embeddings.

3 Experiment

For the research, to investigate how embedding quality varies under the increase of number of topics, we have constructed a controlled experimental framework . The study examines information retrieval effectiveness under varying topic complexity within documents. As web search and information retrieval fundamentally rely on query–document matching, similarity between embeddings is computed using cosine similarity, a standard metric for measuring how different vectors align in high-dimensional vector space.

We have decided to use the nomic-embex-text language model, mainly for its light computational cost and established performance in carrying out such retrieval tasks. For each query–document pair, cosine similarity scores are computed to produce a ranked retrieval list. Retrieval effectiveness is then evaluated against true retrieval result to achieve an accuracy score. Changes in accuracy scores are analysed across these conditions to determine whether performance degradation occurs as the number of topics within documents increases, thereby providing empirical evidence regarding the representational capacity of document embeddings.

[1] SIGIR refers to the International ACM SIGIR Conference on Research and Development in Information Retrieval, a leading annual conference in the field of information retrieval organised by the Association for Computing Machinery (ACM).

3.1 Datasets

The datasets used in this study are accessed through *ir_datasets*, a standardised interface that provides uniform access to a wide range of information retrieval benchmarking collections. The use of a common interface ensures consistency in data structure, formatting and relevance judgements, thereby reducing implementation bias and improving reproducibility.

All datasets retrieved through *ir_datasets* share three essential components relevant to this research:

- **Queries**: Natural language queries issued by users, used to retrieve relevant documents.
- **Documents**: A collection of textual records against which retrieval is performed.
- **Qrels (Query Relevance Judgements)**: Ground-truth annotations specifying which documents are relevant to each query.

Two benchmark datasets with differing characteristics were selected to evaluate the robustness of the experimental findings.

3.1.1 Dataset 1: Cranfield

The Cranfield collection is a classical information retrieval benchmark consisting of 1,400 documents and 225 queries. The documents are technical abstracts in the field of aeronautical engineering, covering topics such as aircraft design, fluid dynamics and propulsion systems. Each query is accompanied by manually curated relevance judgements. The relatively controlled and domain-specific nature of Cranfield makes it well-suited for analyzing retrieval behavior under structured experimental conditions.

3.1.2 Dataset 2: LISA

The LISA (Library and Information Science Abstracts) dataset contains 6,004 documents and 35 queries. The documents consist of abstracts from the field of library and information science, covering topics such as information systems, cataloging, indexing, and information management. Relevance judgments are provided for each query. Compared to Cranfield, LISA is larger in document size but contains fewer queries, offering a contrasting evaluation setting that allows examination of embedding behavior under different corpus characteristics.

3.2 Evaluation Methods

3.2.1 Precision@K

Precision@k measures the proportion of relevant documents among the top k retrieved results (Manning et al. 2008).

$$\text{Precision@}k = \frac{\sum_{i=1}^{k} \text{rel}(i)}{k} \tag{1}$$

Where:

- rel(i) = 1 if the document at rank i is relevant, and 0 otherwise

Precision@K focuses exclusively on the highest-ranked results, evaluating early retrieval accuracy but does not take in to account the relative ordering of relevant documents within the top k.

For our research, we set k as 10, which stems from a real-world user's behavior when surfing the web. For most searches, people tend to only focus on the first 10 returned results; hence, Precision@K measures how many of the top 10 results are actually useful.

3.2.2 Reciprocal Rank

Reciprocal Rank computes the inverse of the rank of the first relevant document (Manning et al. 2008).

$$\mathrm{RR} = \frac{1}{\min\{i \mid \mathrm{rel}(i) = 1\}} \tag{2}$$

Where:

- rel(i) = 1 if the document at rank i is relevant, and 0 otherwise

Reciprocal Rank evaluates how quickly a user encounters the first relevant result. The metric ignores subsequent relevant documents and concentrates solely on the earliest successful match. This benchmark bears great resemblance to when users search for specific pieces of information, such as research papers, a specific product page or an official website. If the correct result is retrieved at rank 1, reciprocal rank score is maximal.

3.2.3 Average Precision

Average Precision evaluates ranking quality by averaging precision values at the ranks where relevant documents occur (Manning et al. 2008).

$$\mathrm{AP} = \frac{1}{R}\sum_{k=1}^{n} \mathrm{Precision@}k \cdot \mathrm{rel}(k) \tag{3}$$

Where:

- R = total number of relevant documents
- n = total retrieved documents
- rel(k) = 1 if the document at rank i is relevant, and 0 otherwise

Average Precision considers not only the number of relevant documents retrieved, but also their respective ranking positions. Earlier relevant documents contribute more to the final score. This metric evaluates overall ranking quality, which is suitable when users inspect multiple relevant documents rather than searching for a single one.

With each metrics reflecting different focuses, together they provide a comprehensive evaluation of how embedding quality changes as the number of topics within a single document increases.

3.3 Experimental Design

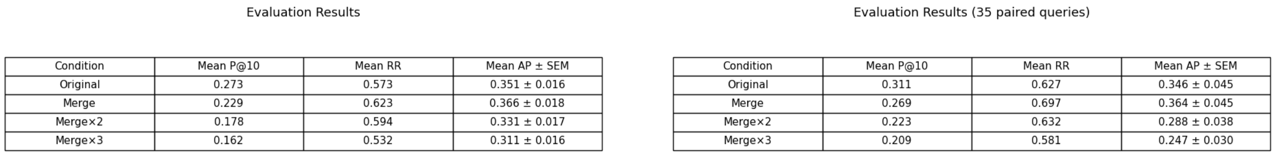
To give a further breakdown into the research process, we have divided the procedure into 6 different steps.

- Step 1: Gather documents where each document only contains one topic
- Step 2: Evaluate the Search accuracy on the documents that only contains one topic
- Step 3: Increase the number of topics that each document contain by merging one by one
- Step 4: Re-evaluate the Web Search's Accuracy
- Step 5: Repeat merging
- Step 6: Observe the change in Web Search's Accuracy

The assessment of Search's Accuracy is conducted under three different evaluation benchmark aforementioned: Precision@K, Reciprocal Rank and Average Precision, while paired t-test is also carried out to test the significance level of our results, thereby validating whether such changes in Accuracy are due to randomness or it actually reflects systematic effects arising from experimental manipulation.

3.4 Findings

This section presents the findings obtained from the merging and evaluation procedure described in the Experimental Design. We begin with the summary results table for the Cranfield dataset, followed by a performance decay curve to visualise trends across embedding conditions. We then examine query-level behaviour through the Average Precision plot with error bars before conducting paired significance tests to determine whether observed differences are statistically reliable. The same analytical sequence is subsequently applied to the LISA dataset to ensure consistency and comparability across datasets. Results from both datasets are presented side-by-side, since we found out that they seemingly follow the same patterns for each designated visualizations of different purposes.



Evaluation Results

Condition	Mean P@10	Mean RR	Mean AP ± SEM
Original	0.273	0.573	0.351 ± 0.016
Merge	0.229	0.623	0.366 ± 0.018
Merge×2	0.178	0.594	0.331 ± 0.017
Merge×3	0.162	0.532	0.311 ± 0.016

(a) Cranfield

Evaluation Results (35 paired queries)

Condition	Mean P@10	Mean RR	Mean AP ± SEM
Original	0.311	0.627	0.346 ± 0.045
Merge	0.269	0.697	0.364 ± 0.045
Merge×2	0.223	0.632	0.288 ± 0.038
Merge×3	0.209	0.581	0.247 ± 0.030

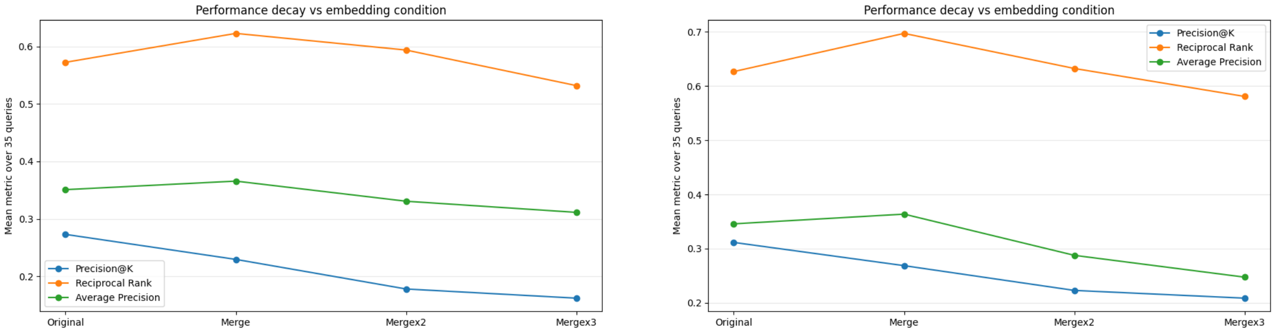
(b) LISA

Figure 1: Evaluation results across embedding conditions for both datasets

The evaluation results for Cranfield and LISA are presented side by side in Figure 1. At first glance, both datasets exhibit a consistent degradation trend as the number of topics encoded within a single document increases. Precision@10 declines monotonically across merging conditions in both corpora. For Cranfield, it decreases from 0.273 in the Original condition to 0.162 under Merge×3. A similar pattern is observed in LISA, where Precision@10 drops from 0.311 to 0.209. This consistent reduction indicates that early retrieval accuracy is highly sensitive to increasing semantic complexity.

A slightly different tendency could be seen for Average Precision and Reciprocal Rank. Specifically, such scores experience a marginal growth when moving from Original to Merge condition, but under more intense merging impacts, both metrics decline progressively, giving indications that the prior increase might only be a temporary phenomenon rather than a representative trend.

To understand the overall trajectory of the accuracy decline, we have to inspect the performance decay curve across three evaluation benchmarks.



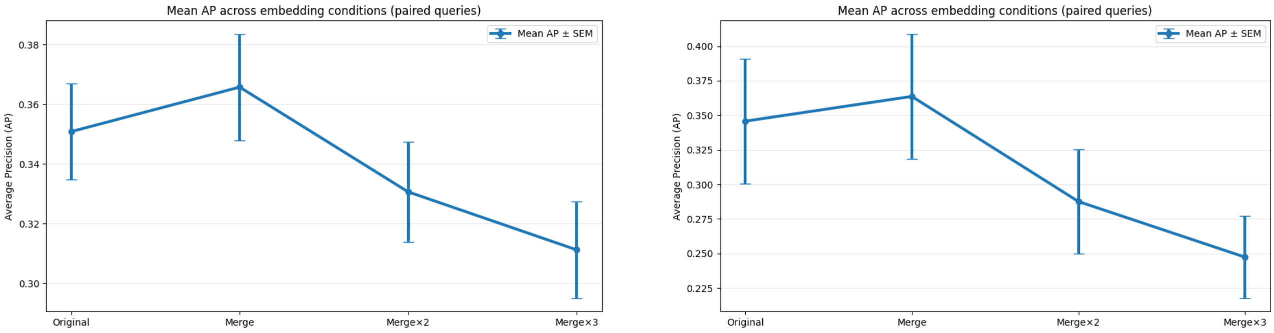
(a) Cranfield

(b) LISA

Figure 2: Performance decay curves across embedding conditions

In both datasets, the blue curve illustrates that the steepest decrease lies in Precision@10, suggesting that early retrieval quality is particularly sensitive to the change in topic complexity. Average precision appears to be declining more gradually while reciprocal rank remains fairly stable at the first merge; nevertheless, they eventually decrease under heavier compression.

Average Precision plots with Standard Error in Means offer a closer look into query-level variability.



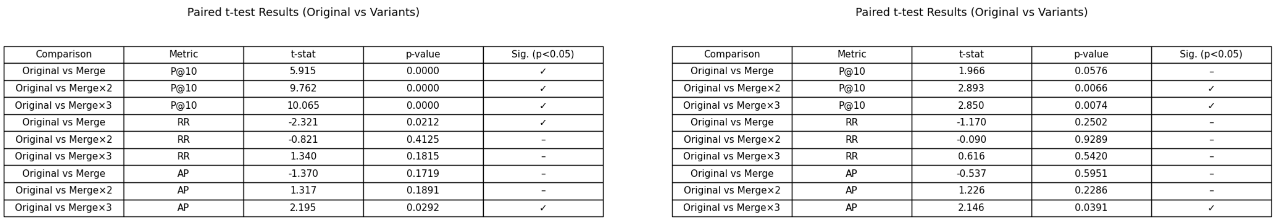
(a) Cranfield

(b) LISA

Figure 3: Mean Average Precision with standard error

Regarding Cranfield, small error margins show that degradtion is not due by chance or affected by outliers, but rather an outcome of systematic changes. LISA exhibits slightly higher variability, but this range is exceeded by the downward shift in mean Average Precision under Mergex2 and Mergex3, once again reinforcing the interpretation of statistically meaningful degradation.

Finally, paired t-tests were conducted to determine whether the observed performance differences across embedding conditions are statistically reliable.



Paired t-test Results (Original vs Variants)

Comparison	Metric	t-stat	p-value	Sig. (p<0.05)
Original vs Merge	P@10	5.915	0.0000	✓
Original vs Merge×2	P@10	9.762	0.0000	✓
Original vs Merge×3	P@10	10.065	0.0000	✓
Original vs Merge	RR	-2.321	0.0212	✓
Original vs Merge×2	RR	-0.821	0.4125	–
Original vs Merge×3	RR	1.340	0.1815	–
Original vs Merge	AP	-1.370	0.1719	–
Original vs Merge×2	AP	1.317	0.1891	–
Original vs Merge×3	AP	2.195	0.0292	✓

(a) Cranfield

Paired t-test Results (Original vs Variants)

Comparison	Metric	t-stat	p-value	Sig. (p<0.05)
Original vs Merge	P@10	1.966	0.0576	–
Original vs Merge×2	P@10	2.893	0.0066	✓
Original vs Merge×3	P@10	2.850	0.0074	✓
Original vs Merge	RR	-1.170	0.2502	–
Original vs Merge×2	RR	-0.090	0.9289	–
Original vs Merge×3	RR	0.616	0.5420	–
Original vs Merge	AP	-0.537	0.5951	–
Original vs Merge×2	AP	1.226	0.2286	–
Original vs Merge×3	AP	2.146	0.0391	✓

(b) LISA

Figure 4: Paired t-test results comparing Original embeddings against merged conditions.

The results in Figure 4 indicate that Precision@10 is the most consistently affected metric across both datasets. In Cranfield, decreases in Precision@10 are statistically significant under all merging conditions (all p-values ≈ 0.0000), confirming a systematic decline in early retrieval accuracy as topic multiplicity increases. In LISA, significance emerges under stronger merging (Merge×2 and Merge×3), while the Original vs Merge comparison remains non-significant ($p = 0.0576$), suggesting that light semantic aggregation does not immediately impair top-ranked precision but degradation becomes detectable beyond a moderate threshold.

For Reciprocal Rank and Average Precision, significance appears more selectively. Reciprocal Rank shows largely non-significant differences across conditions, indicating that the rank of the first relevant document remains relatively stable under semantic compression. Average Precision becomes statistically significant only under the highest topic multiplicity (Merge×3) for both datasets, suggesting that overall ranking quality deteriorates progressively rather than abruptly. Importantly, the presence of both significant and non-significant outcomes strengthens the validity of the findings, demonstrating that degradation is metric-sensitive and systematic rather than a result of experimental bias. More work would be done in the future to address this prompt.

4 Conclusion

In conclusion, the experiment results has led to two main important findings:

- According to Precision@K benchmark, the capacity is one topic.
- According to Reciprocal Rank and Average Precision benchmarks, the capacity is 2 topics.

While it remains to be further investigated, our intuition is that a single embedding can encode multiple topics, but these topics compete for representational and ranking priority within a fixed-dimensional vector. When examining the mathematical definitions of each benchmark, this behaviour becomes clearer. Precision@10 applies a strict cutoff at rank k; therefore, even a small ranking shift caused by topic compression may push a previously relevant document outside the top 10, resulting in an immediate and visible decline in accuracy after the first merging step.

In contrast, Reciprocal Rank and Average Precision are more rank-sensitive rather than cutoff-based. Reciprocal Rank depends only on the position of the first relevant document, meaning that unless that specific document is displaced significantly, the metric remains relatively stable. Average Precision further distributes weight across all relevant documents in the ranked list, allowing moderate rank shifts without immediate collapse in score. Consequently, while semantic compression introduces interference between topics, its impact manifests gradually in rank-aware metrics, becoming statistically evident only when topic multiplicity exceeds the embedding's effective representational capacity.

5 Discussion

The research suggests that single-vector document embeddings hold inherent shortcomings, especially when the semantic complexity accumulates. While moderate topic aggregation might not instantly affect retrieval accuracy, there is a definite threshold to when this degradation becomes systematic.

For future considerations, we aim at investigating whether multi-vector embedding approaches can alleviate this capacity constraint. For example, QDER has introduced a query-specific multi-vector framework that preserves token- and entity-level representations and performs late aggregation only at the final scoring stage, reducing semantic compression within a single fixed vector (Chatterjee & Dalton 2025). In the future, a direct comparative study between single-vector and multi-vector embeddings would provide valuable insight into whether representational capability could be extended through new technical modifications.

References

Aghaei, S. et al. (2022), 'Interactive search on the web: The story so far', *Information* **13**(7), 324.

Chatterjee, S. & Dalton, J. (2025), Qder: Query-specific document and entity representations for multi-vector document re-ranking, *in* 'Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)'.
URL: *https://dl.acm.org/doi/10.1145/3726302.3730065*

Gómez, J. & Vázquez, P.-P. (2022), 'An empirical evaluation of document embeddings and similarity metrics for scientific articles', *Applied Sciences* **12**(11), 5664.

Kesiraju, S., Plchot, O., Burget, L. & Gangashetty, S. V. (2020), 'Learning document embeddings along with their uncertainties', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* .
URL: *https://dl.acm.org/doi/10.1109/TASLP.2020.3012062*

Khattab, O. & Zaharia, M. (2020), Colbert: Efficient and effective passage search via contextualized late interaction over bert, *in* 'Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)'.
URL: *https://dl.acm.org/doi/10.1145/3397271.3401075*

Manning, C. D., Raghavan, P. & Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press.

Veluru, S. R., Marella, V. C. & Erukude, S. T. (2025), 'The evolution of search engines: From keyword matching to ai-powered understanding', *International Journal of Computer Applications* **187**(18), 7–14.