 DOI: https://doi.org/10.26493/scores26.18

Geometric Characterization of Embedding Spaces Across Chunking Strategies and Embedding Models for Slovenian Text

Anja Jodl
anja.jodl@student.um.si
Faculty of Electrical Engineering and Computer Science,
University of Maribor
SI-2000 Maribor, Slovenia

Mitko Nikov
mitko.nikov@um.si
Faculty of Electrical Engineering and Computer Science,
University of Maribor
SI-2000 Maribor, Slovenia

ABSTRACT

This paper investigates whether geometric properties of embedding spaces can provide useful intrinsic indicators for evaluating text chunking strategies and embedding models without relying on downstream tasks. We evaluate several text chunking strategies with different overlap configurations using five embedding models on a large collection of historical Slovenian newspapers. Our results show that embedding-space geometry is strongly model-dependent, while chunking strategies substantially affect local continuity, cluster separation, and how well cluster centroids represent the original article. Token-based chunking with approximately 256 tokens and 25% overlap provided the most consistent balance between the evaluated geometric properties and computational requirements. The results also show that no individual geometric metric defines embedding-space quality and that complementary properties should therefore be interpreted jointly.

KEYWORDS

text embeddings, document chunking, embedding space geometry, semantic similarity, low-resource languages, retrieval, historical documents

1 INTRODUCTION

Dense retrieval systems convert sections of text, so-called chunks, into vectors using embedding models, and retrieve content whose vector representations are similar to the representation of the user's query. In retrieval-augmented generation (RAG), these retrieved text sections provide context for generating accurate answers. The choice of where to split the text, also known as the chunking strategy, determines which text is represented together, how much content is repeated across chunks, and how many representations must be stored and processed. Since the resulting chunks are encoded via an embedding model, chunking strategies and embedding models are interdependent and their selection should be considered together when designing a retrieval pipeline [11].

Chunking strategies can be evaluated through downstream retrieval or task performance. Although such evaluations are essential for establishing effectiveness in a particular application, their results reflect the combined effects of multiple components, including chunking, embedding models, retrieval algorithms, and reranking methods [14, 19]. Final performance scores alone therefore provide limited insight into how chunking and embedding choices affect the underlying vector representations.

When task-specific queries and relevance judgements are unavailable, an additional challenge is how to characterize these effects without first constructing a downstream evaluation dataset. Therefore, previous work has used geometric properties to analyze embedding quality [4, 17], while others applied a cluster-validity measure over sentence embeddings for reference-free text segmentation evaluation [8].

This research investigates how chunking strategies and embedding models affect local continuity between consecutive chunks, article-level separation, the global distribution of embedding variance, and agreement between chunk-based and unchunked article representations. The objective is to characterize these geometric properties jointly and examine where they provide consistent or conflicting indications about representation behavior. They are treated as intrinsic diagnostics rather than as validated predictors of downstream retrieval performance.

Hence, the contributions of this work are:

- A geometric evaluation protocol combining complementary measures of continuity, clustering, variance, and representation preservation.
- Operational definitions of semantic drop rate (SDR) and mean representation degradation (MRD).
- An empirical comparison of 17 chunking configurations and five embedding models, accompanied by publicly available code and experimental results [5].

2 METHODOLOGY

To evaluate how chunking strategies and embedding models affect the resulting embedding spaces, a multi-stage experimental pipeline was constructed. The pipeline consists of OCR digitization of archived newspapers; statistical analysis of the extracted articles; filtering and cleaning the articles based on their text quality and statistical outliers; chunking them using various strategies; generating embeddings with different models; and computing geometric metrics to evaluate the resulting embedding spaces.

2.1 Dataset

To compile the machine-readable dataset, 3430 scanned Slovenian-language newspapers published between 1900 and 2015 and provided by the Digital Library of Slovenia [10] were processed using MinerU [15] and the PaddleOCR engine [2]. The extracted content was stored in Markdown format.

Individual articles were isolated by splitting newspaper files according to Markdown header structures. In total, 167 977 articles were extracted, of which 90 307 were retained after cleaning and filtering.

The cleaning procedure removed non-text elements introduced during extraction, duplicate articles, and articles failing text-quality requirements based on length, letter and vowel ratios, and average

 DOI: https://doi.org/10.26493/scores26.18

word length. Additional filtering removed articles with excessive numeric or special characters and statistical outliers in text-structure measures using median absolute deviation-based bounds. The complete filtering configuration and implementation are publicly available in the accompanying repository [5].

Most articles are written in historical standard Slovenian, which differs from contemporary Slovenian in vocabulary, grammar, and orthography. This variation presents an additional challenge for embedding models typically trained on modern language corpora.

2.2 Chunking Strategies

Following extraction and cleaning, the articles were processed using the chunking strategies shown in Table 1. These included no-operation (no-op) chunking, in which the complete article was retained as a single text segment; paragraph chunking without overlap; sentence-based chunking using groups of N sentences with an overlap of M sentences (SN/M); token-based chunking using chunks of N tokens with an overlap of M tokens (TN/M); and recursive chunking using the LangChain recursive splitter [6], with chunks of N characters and an overlap of M characters (RN/M).

Table 1: Chunking strategy statistics.

Chunking	Chunks	C/A	Ch/C	Memory	Overlap
No-op	90,307	1.00	2471.59	213 MB	–
P	750,276	8.31	295.69	212 MB	0.0%
S1/0	2,641,788	29.25	83.21	210 MB	0.2%
S4/0	694,511	7.69	319.33	212 MB	0.0%
S4/1	880,608	9.75	326.66	274 MB	26.2%
T32/0	2,346,500	25.98	95.52	214 MB	0.0%
T128/0	619,968	6.87	360.65	213 MB	0.0%
T128/16	689,322	7.63	366.55	241 MB	14.9%
T256/0	330,820	3.66	675.61	213 MB	0.0%
T256/32	359,251	3.98	694.67	238 MB	17.8%
T256/64	397,002	4.40	712.65	270 MB	30.6%
T256/128	529,784	5.87	743.24	376 MB	53.4%
T256/192	930,188	10.30	764.82	678 MB	76.0%
R400/0	778,317	8.62	284.99	212 MB	0.0%
R400/64	804,092	8.90	298.73	229 MB	12.7%
R1000/0	317,105	3.51	700.96	212 MB	0.0%
R1000/100	319,127	3.53	710.34	216 MB	5.4%

C/A = mean chunks per article, Ch/C = mean characters per chunk, Overlap = mean overlap between consecutive chunks

Table 1 reports the number of generated chunks, mean chunks per article (C/A), mean characters per chunk (Ch/C), and the observed mean overlap between consecutive chunks. The latter may differ from the configured overlap due to article boundaries.

Figure 1 shows the resulting chunk-length distributions and indicates that paragraph- and sentence-based strategies produce more variable chunk sizes, whereas token-based strategies produce lengths more concentrated around their configured sizes.

T32/0 was included as an extreme small-chunk configuration to assess whether the evaluated geometric properties distinguish it from configurations using longer contexts. No-op chunking provides an unchunked reference representation of each article and is used as a baseline for metrics comparing chunk-based and complete-article representations.

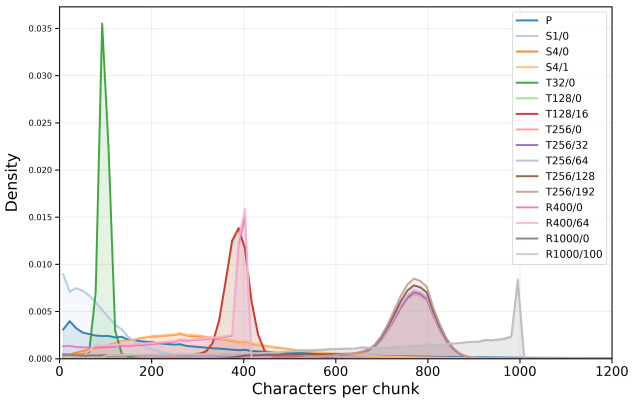


Figure 1: Probability densities of chunk character counts for different chunking strategies. The area under each distribution integrates to 1.

2.3 Embedding Models

Five open-source embedding models were selected to evaluate whether the effects of chunking remain consistent across different embedding spaces, as shown in Table 2: Microsoft's Multilingual E5 Large (E5) [16], Google's 300-million-parameter EmbeddingGemma [13], Alibaba Cloud's 600-million-parameter Qwen3 Embedding model [18], BAAI's BGE-M3 [1], and Novosel's SloVlo V1 [9], which is based on Microsoft's Multilingual E5 Base model. No instruction prefixes were used for any of the embedding models.

Table 2: Architectures of the evaluated embedding models.

Model	Released	Params.	Layers	Dim.
Multilingual E5 Large	Jun. 2023	600M	24	1024
SloVlo V1	Jun. 2024	300M	12	768
BGE-M3	Jan. 2024	560M	24	1024
Qwen3 Embedding 0.6B	Jun. 2025	600M	28	1024
EmbeddingGemma	Sep. 2025	300M	24	768

Params = parameters, Dim = embedding dimensions

2.4 Geometric Evaluation

After generating embeddings for each combination of chunking strategy and embedding model, the resulting embedding spaces were evaluated through complementary geometric properties.

A meaningful embedding space should preserve semantic relationships by placing related content close together while maintaining separation between distinct content. For chunked articles, this includes preserving continuity between consecutive chunks, separating content from different articles, avoiding excessive concentration along particular directions of the embedding space, and preserving article-level information when an article is represented through its chunks [4, 7, 12].

The evaluation therefore examined local semantic continuity, separation between articles, the global distribution of embeddings, and the preservation of article-level representations. All embeddings were normalized to unit length before evaluation to ensure consistent cosine similarity comparisons.

3 RESULTS

The complete pipeline was executed on a workstation equipped with an Intel i9-13900K CPU, an NVIDIA RTX 4090 with 24 GB of VRAM, and 64 GB of RAM. Article extraction and cleaning required approximately 30 minutes, application of all chunking strategies close to 4.5 hours, and embedding generation approximately 124 hours. Embedding generation therefore accounted for the majority of the computational cost. Results and compute metadata were stored in a SQLite database of approximately 320 GB. All code and complete results are available in the accompanying repository [5].

As an initial comparison, pairwise cosine similarity distributions were examined for random chunk pairs from different articles (Negative), chunk pairs from the same article (Positive), and consecutive chunks from the same article (Transition). Mean negative cosine similarity (MNCS), mean positive cosine similarity (MPCS), and mean transition cosine similarity (MTCS) are reported for selected configurations in Table 3. MNCS and MPCS describe article-level separability, while MTCS reflects local continuity.

Figure 2 illustrates the distributions for Gemma T256/64 and E5 T256/64. The distributions differ substantially between models, demonstrating that cosine similarity values should be interpreted relative to the geometry of each embedding space rather than compared directly across models.

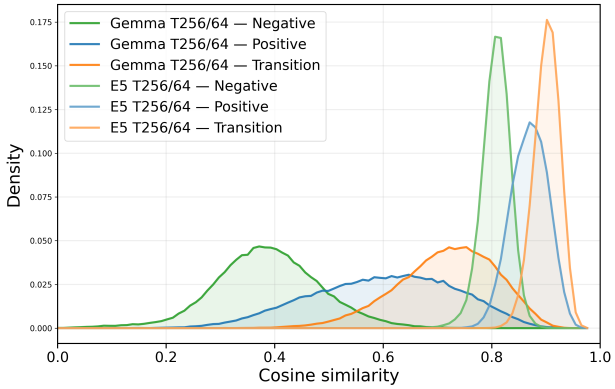


Figure 2: Pairwise cosine similarity distributions for Gemma T256/64 and E5 T256/64 embeddings, shown for negative, positive, and transition chunk pairs.

High MTCS alone does not necessarily indicate better representations, as local continuity should be interpreted relative to the similarity of unrelated content. The results show substantial differences between embedding models as the older E5 and SloVlo models exhibit higher cosine similarity values across all three metrics than BGE, Qwen3, and Gemma, indicating more compressed embedding spaces with less separation between articles.

To further evaluate the distinction between different types of chunk pairs, a Gaussian Naive Bayes (GNB) classifier was trained for each combination of embedding model and chunking strategy using 100 000 randomly sampled pairs per class. Cosine similarity was used as the only feature, with 80% of the pairs used for training and 20% for testing. Classification was performed between positive and negative pairs (N/P) and between negative and transition pairs (N/T).

Classification accuracy depends strongly on both the embedding model and chunking strategy. Accuracy was consistently higher for distinguishing negative from transition pairs, which is expected because consecutive chunks naturally share semantic and linguistic context. Although E5 and SloVlo exhibit higher overall cosine similarities, they achieved higher overall classification accuracy than BGE, Qwen3, and Gemma on this dataset.

To quantify local continuity, semantic drop rate (SDR) was defined as the proportion of consecutive chunk pairs whose cosine similarity falls below the mean negative cosine similarity (MNCS) of the corresponding model and chunking strategy. Lower SDR indicates better preservation of continuity.

Overlapping strategies substantially reduce SDR compared with non-overlapping strategies, as shared content increases the similarity between consecutive chunks. Even relatively small overlaps can produce considerable improvements, as observed for R1000/100.

The Davies-Bouldin index (DBI) [3] was computed to evaluate article-level cluster separation, with lower values indicating better separation. Token-based strategies, particularly T256 with overlap, consistently achieve lower DBI values than the other strategies, indicating more compact and better-separated article clusters.

Mean representation degradation (MRD) was used to measure how closely the centroid of an article's chunk embeddings corresponds to the embedding of the original article. MRD is defined as the mean cosine distance between the original article embedding and the corresponding chunk centroid. Because cosine similarity distributions differ between models, MRD is used only for comparing chunking strategies within the same model. In most cases, token-based strategies, particularly T256 with overlap, achieve lower MRD values than recursive chunking.

Finally, principal component analysis (PCA) was used to examine the global distribution of variance in the embedding spaces. The cumulative variance ratios of the first 1, 5, and 10 principal components are reported as PC1, PC5, and PC10 in Table 3. No clear patterns were observed among chunking strategies or consistent differences between models.

4 CONCLUSION

This research demonstrates that geometric properties of embedding spaces can provide diagnostic intrinsic indicators for evaluating chunking strategies and embedding models when task-specific queries and relevance judgements are unavailable.

The results show that embedding-space geometry is strongly model-dependent, while chunking strategies substantially affect local semantic continuity, article separation, and the preservation of article-level representations. No individual metric defines embedding-space quality, therefore the evaluated properties should be interpreted jointly.

Token-based chunking with moderate overlap provided the most consistent balance across the evaluated metrics. Even small overlaps produced substantial improvements, while overlap above 25% generally resulted in diminishing benefits relative to the increased number of chunks and computational requirements. Among the evaluated configurations, Qwen3 T256/64 and Gemma T256/64 provided a strong overall balance, with Gemma achieving comparable results using approximately half as many parameters.

Table 3: Selected set of the evaluation results combined by chunking strategy and embedding model.

Model	Chunking	MNCS ↓	MPCS↑	MTCS ↑	SDR ↓	DBI ↓	MRD ↓	PC1 ↓	PC5 ↓	PC10 ↓	GNB N/P ↑	GNB N/T ↑
E5	P	0.7770	0.8154	0.8286	10.44%	4.5847	0.0529	0.0450	0.1156	0.1668	72.60%	78.10%
	S1/0	0.7808	0.8002	0.8171	15.49%	8.1291	0.0745	0.0585	0.1262	0.1803	62.44%	72.16%
	S4/1	0.7854	0.8307	0.8811	0.62%	4.2962	0.0509	0.0299	0.1025	0.1548	76.57%	94.55%
	T32/0	0.8091	0.8286	0.8445	7.54%	7.7732	0.0796	0.0294	0.0993	0.1511	65.59%	77.09%
	T256/0	0.8041	0.8605	0.8729	3.83%	3.2015	0.0471	0.0330	0.1172	0.1751	82.87%	88.27%
	T256/64	0.8078	0.8692	0.9010	0.04%	3.1413	0.0455	0.0348	0.1146	0.1735	84.42%	97.01%
	R1000/100	0.7933	0.8461	0.8625	3.56%	3.2629	0.0485	0.0315	0.1093	0.1656	79.45%	86.95%
SloVlo	P	0.6668	0.7383	0.7553	9.70%	4.5019	0.0895	0.0501	0.1518	0.2203	74.18%	78.17%
	S1/0	0.7052	0.7381	0.7617	18.28%	7.9990	0.1315	0.0800	0.1849	0.2541	61.58%	69.62%
	S4/1	0.6617	0.7396	0.8218	0.52%	4.2035	0.0857	0.0349	0.1314	0.2024	75.64%	93.16%
	T32/0	0.7254	0.7576	0.7820	9.93%	7.7314	0.1525	0.0410	0.1395	0.2114	64.53%	74.10%
	T256/0	0.6718	0.7723	0.7970	2.68%	3.1663	0.0845	0.0424	0.1393	0.2113	82.04%	88.52%
	T256/64	0.6740	0.7851	0.8431	0.03%	3.0946	0.0826	0.0444	0.1386	0.2109	83.22%	96.19%
	R1000/100	0.6610	0.7593	0.7901	3.41%	3.1964	0.0823	0.0423	0.1362	0.2075	80.20%	87.53%
BGE	P	0.3461	0.4442	0.4916	11.16%	4.6069	0.1597	0.0289	0.1074	0.1655	72.46%	79.08%
	S1/0	0.3624	0.4168	0.4701	13.68%	8.0167	0.2127	0.0402	0.1191	0.1805	63.07%	74.33%
	S4/1	0.3667	0.4900	0.6548	0.52%	4.2885	0.1601	0.0306	0.1082	0.1675	75.77%	95.05%
	T32/0	0.3723	0.4342	0.4974	6.19%	7.5657	0.2277	0.0246	0.0989	0.1620	66.53%	79.94%
	T256/0	0.3854	0.5443	0.5950	2.35%	3.2251	0.1721	0.0315	0.1140	0.1738	82.80%	90.24%
	T256/64	0.3907	0.5698	0.6949	0.02%	3.1278	0.1684	0.0328	0.1169	0.1768	83.87%	98.00%
	R1000/100	0.3741	0.5255	0.5824	3.64%	3.2920	0.1700	0.0313	0.1142	0.1736	80.17%	88.37%
Qwen3	P	0.3179	0.4060	0.4572	15.19%	4.8259	0.1274	0.0398	0.1321	0.1973	66.83%	74.08%
	S1/0	0.2941	0.3409	0.3835	22.61%	8.9303	0.2005	0.0555	0.1361	0.1954	59.60%	67.30%
	S4/1	0.3514	0.4658	0.6171	2.28%	4.5609	0.1318	0.0367	0.1213	0.1863	71.58%	90.42%
	T32/0	0.3348	0.3864	0.4275	14.59%	8.5227	0.2099	0.0291	0.0986	0.1556	61.93%	70.36%
	T256/0	0.3718	0.5303	0.5805	4.50%	3.3189	0.1110	0.0382	0.1280	0.1951	78.92%	86.47%
	T256/64	0.3784	0.5559	0.6771	0.22%	3.2401	0.1135	0.0396	0.1310	0.1982	80.43%	95.33%
	R1000/100	0.3648	0.5161	0.5740	4.84%	3.3719	0.1067	0.0386	0.1305	0.1983	76.02%	85.10%
Gemma	P	0.3445	0.4783	0.5034	13.33%	4.7226	0.1206	0.0644	0.1675	0.2313	71.29%	74.41%
	S1/0	0.4068	0.4670	0.4936	23.88%	8.6883	0.2461	0.0716	0.1657	0.2277	60.00%	64.49%
	S4/1	0.3556	0.4970	0.6494	1.49%	4.3987	0.1179	0.0395	0.1331	0.2004	72.76%	90.69%
	T32/0	0.4473	0.5087	0.5410	18.07%	8.2395	0.2732	0.0359	0.1183	0.1820	61.36%	67.19%
	T256/0	0.3773	0.5691	0.6127	4.51%	3.2290	0.0768	0.0451	0.1495	0.2202	80.38%	86.30%
	T256/64	0.3862	0.5989	0.7155	0.09%	3.1156	0.0796	0.0477	0.1514	0.2204	81.53%	95.24%
	R1000/100	0.3794	0.5505	0.6033	5.41%	3.3065	0.0752	0.0429	0.1427	0.2096	76.20%	83.80%

5 FUTURE WORK

Future work should validate the relationship between the proposed geometric metrics and downstream retrieval performance on datasets containing queries and relevance judgements. Additionally, the geometric metrics developed in this research could potentially be used not only for evaluation but also to guide the chunking process itself. That includes investigating adaptive chunking algorithms that determine chunk boundaries by optimizing intrinsic geometric properties such as intra-chunk cohesion, anisotropy, or neighborhood preservation.

6 ACKNOWLEDGMENTS

The research paper was elaborated within the research project no. J7-60128: AID HCH - Breakthrough in humanities and cultural heritage with artificial intelligence, funded by the Slovenian Research and Innovation Agency (ARIS).

REFERENCES

[1] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216 [cs.CL]

[2] Cheng Cui et al. 2025. PaddleOCR 3.0 Technical Report. arXiv:2507.05595 [cs.CV] https://arxiv.org/abs/2507.05595

[3] David L. Davies and Donald W. Bouldin. 1979. A Cluster Separation Measure. PAMI-1, 2 (1979), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

[4] Kawin Ethayarajh. 2019. *How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings.* https://doi.org/10.48550/arXiv.1909.00512 arXiv:1909.00512 [cs.CL]

[5] Anja Jodl and Mitko Nikov. 2026. Git Repository for Geometric Characterization of Embedding Spaces. https://github.com/GeMMA-FERI/embedding-space-characterization

[6] LangChain. 2026. Documentation for RecursiveCharacterTextSplitter. https://reference.langchain.com/. Accessed 26 August 2026.

[7] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. *On the Sentence Embeddings from Pre-trained Language Models.* https://doi.org/10.48550/arXiv.2011.05864 arXiv:2011.05864 [cs.CL]

[8] Evan Lucas, Dylan Kangas, and Timothy Havens. 2023. A Reference-Free Segmentation Quality Index (SegReFree). In *Findings of the Association for Computational Linguistics: EMNLP 2023.* Association for Computational Linguistics, Singapore, 2957–2968. https://doi.org/10.18653/v1/2023.findings-emnlp.195

[9] Rok Novosel. 2024. SloVlo V1: Embeddings for the Slovenian Language. https://rokn.io/posts/slovlo-v1-embeddings-for-the-slovenian-language

[10] NUK. 2026. Digital Library of Slovenia. https://www.dlib.si/

[11] Renyi Qu, Ruixuan Tu, and Forrest Bao. 2025. Is Semantic Chunking Worth the Computational Cost?. In *Findings of the Association for Computational Linguistics: NAACL 2025.* Association for Computational Linguistics, 2155–2177. https://doi.org/10.18653/v1/2025.findings-naacl.114

[12] Sara Rajaee and Mohammad Taher Pilehvar. 2022. An Isotropy Analysis in the Multilingual BERT Embedding Space. In *Findings of the Association for Computational Linguistics: ACL 2022* (Dublin, Ireland, 2022-05). Association for Computational Linguistics, 1309–1316. https://doi.org/10.18653/v1/2022.findings-acl.103

[13] Henrique Schechter Vera, Sahil Dua, et al. 2025. EmbeddingGemma: Powerful and Lightweight Text Representations. (2025). https://arxiv.org/abs/2509.20354

[14] Muhammad Arslan Shaukat, Muntasir Adnan, and Carlos C. N. Kuhn. 2026. *A Systematic Investigation of Document Chunking Strategies and Embedding Sensitivity.* https://doi.org/10.48550/arXiv.2603.06976 arXiv:2603.06976 [cs.CL]

[15] Bin Wang et al. 2026. MinerU2.5-Pro: Pushing the Limits of Data-Centric Document Parsing at Scale. *arXiv preprint arXiv:2604.04771* (2026).

[16] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. https://doi.org/10.48550/arXiv.2402.05672 arXiv:2402.05672 [cs.CL]

[17] Brendan Whitaker, Denis Newman-Griffis, Aparajita Haldar, Hakan Ferhatosmanoglu, and Eric Fosler-Lussier. 2019. Characterizing the Impact of Geometric Properties of Word Embeddings on Task Performance. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP.* Association for Computational Linguistics, Minneapolis, USA, 8–17. https://doi.org/10.18653/v1/W19-2002

[18] Yanzhao Zhang et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[19] Yongjie Zhou, Shuai Wang, Bevan Koopman, and Guido Zuccon. 2026. *Beyond Chunk-Then-Embed: A Comprehensive Taxonomy and Evaluation of Document Chunking Strategies for Information Retrieval.* https://doi.org/10.48550/arXiv.2602.16974 arXiv:2602.16974 [cs.IR]