A Comparative Analysis of Multilingual and Monolingual Models for Nepali Legal Document Retrieval

Anjan Parajuli
Independent Researcher
Pokhara, Nepal
anjanparajuli2001@gmail.com

Abstract*—While extensive research has been conducted on information retrieval for high-resource languages, the Nepali language, particularly the Nepali legal domain, remains underexplored. This study aims to address this gap by empirically comparing the performance of multilingual and monolingual open-source language models on a Nepali legal document retrieval task. We constructed a domain-specific dataset consisting of 10 Nepali legal documents. Additionally, 50 curated legal queries were created, with five derived from each document. We evaluated seven multilingual models selected based on their robust performance on the Massive Text Embedding Benchmark (MTEB)[1], alongside three Nepali-specific monolingual models trained exclusively on the Nepali language. The models were evaluated using varying chunk sizes and standard information retrieval metrics, including Recall, Precision, and Mean Reciprocal Rank (MRR). Experimental results demonstrate that the multilingual model BAAI/bge-m3 consistently outperforms the other evaluated models across all settings, achieving 0.92 Recall@6, 0.74 Precision@1, and 0.83 MRR@4. While multilingual models show strong retrieval effectiveness, the findings indicate that existing Nepali monolingual models remain less competitive and require substantial improvement for domain-specific legal retrieval tasks.

Keywords*—*Nepali Language, Legal Information Retrieval, Text Embedding, Multilingual Models, Monolingual Models, Low-Resource Languages

I. Introduction

Information Retrieval (IR) is the task of identifying and retrieving information system resources that are relevant to a user’s information need[2]. IR systems have been extensively studied and applied across a wide range of languages and application domains[3]. With recent advances in artificial intelligence, particularly in machine learning and deep learning, numerous monolingual models (e.g., BERT-based retrievers[4]) and multilingual models (e.g., mBERT, XLM-R[5]) have been developed to improve the effectiveness and efficiency of IR systems. Legal information retrieval and extraction represent an important application area of IR, supporting access to statutes, case laws, and legal documents, and enabling broader dissemination of legal information beyond legal professionals.

While several studies have been conducted on information retrieval for high-resource languages such as English[6] and Chinese[7], the Nepali language remains underexplored due to the lack of large-scale, high-quality annotated datasets. Nepali is a morphologically rich language with significant lexical and syntactic ambiguities[8], which pose challenges for effective text representation and retrieval. These challenges are further intensified in the legal domain, where documents contain domain-specific terminology, formal structures, and complex semantic relationships. Consequently, the effectiveness of existing information retrieval models, particularly monolingual and multilingual approaches, has not been sufficiently investigated for Nepali legal document retrieval.

To address these challenges, this paper presents a comparative analysis of monolingual and multilingual information retrieval models for Nepali legal document retrieval. We first construct a baseline dataset consisting of legal documents that are segmented into structured and semantically meaningful legal chunks. In addition, a query–relevance benchmark is developed to evaluate retrieval performance across different models. Using this setup, we conduct experiments on a range of monolingual and multilingual retrieval models to analyze their effectiveness in the Nepali legal domain. The findings provide insights into the relative strengths of multilingual and monolingual approaches for low-resource legal information retrieval.

The contributions of this paper are threefold. First, we introduce a baseline dataset for Nepali legal document retrieval by constructing structured and semantically meaningful legal chunks from raw legal texts. Second, we develop a query–relevance benchmark to systematically evaluate retrieval performance in the Nepali legal domain. Third, we conduct a comparative analysis of monolingual and multilingual retrieval models, providing empirical insights into their effectiveness for low-resource legal information retrieval.

II. Literature review

Information Retrieval has long been studied using a wide range of techniques, evolving from early Boolean and lexical approaches to modern neural embedding based models[3]. To provide a structured overview of prior work, we categorize existing studies according to the retrieval techniques employed and the languages in which they are applied.

A. Techniques

Information retrieval techniques range from basic logical matching to advanced semantic and neural approaches. The Boolean retrieval model[9], developed in the 1950s, views documents and queries as term sets and applies logical operators (AND, OR, NOT) to identify matching documents. The transition to ranked retrieval began in the 1970s with the Vector Space Model (VSM)[3], which represents documents and queries as weighted term vectors and uses cosine similarity for ranking, with the SMART system at Cornell demonstrating the utility of this paradigm. Simultaneously, probabilistic models such as the probabilistic relevance model provided a principled framework for estimating the likelihood of document relevance and led to the widely used BM25 ranking function[10]. To address semantic structure and term relationships, latent semantic methods such as Latent Semantic Analysis (LSA)[11] were introduced in the late 1980s, using singular value decomposition to capture latent topics in text. The 2000s saw learning-based and graph-based ranking methods such as PageRank[12], which leveraged link structure to improve web search rankings. In recent years, neural and deep learning techniques have transformed IR. Word embedding like Word2Vec[13] enabled semantic similarity at the word level, and transformer-based models such as BERT[4] provide contextualized representations that substantially improve semantic retrieval and re-ranking performance, inspiring models such as ColBERT[14] for efficient passage search.

B. *Languages*

As a frontier in computer systems, English is the most extensively studied language for information retrieval, with most benchmarks and datasets developed for it. Prominent examples include the TREC collections[15], which provide large-scale corpora and relevance judgments for ad hoc retrieval, question answering, and web search tasks. Datasets such as SQuAD[16] and MS MARCO[17] are also widely used and have been translated into multiple languages for multilingual training and evaluation.

Following English, languages such as Chinese[18] and French[19] have received noticeable research attention. However, much of the work in these languages builds upon methodologies, datasets, and evaluation practices originally developed for English. As a result, English remains the primary reference language for advances in retrieval models.

In recent years, information retrieval systems have increasingly adopted multilingual approaches, as large pre-trained models can learn shared linguistic representations and generalize effectively across languages, including those with limited annotated resources.

C. *Nepali Legal Information Retrieval*

Referring to Nepali language, very limited research has been conducted on information retrieval in general. Existing studies include question answering–based retrieval over biomedical research papers[20] and knowledge graph–based retrieval using Wiki-data[21], representing some of the few efforts in Nepali information retrieval.

In the legal domain, research is even more scarce. Enhanced Retrieval for QA Systems Tailored for Nepali Legal Documents Focusing on PSC Examinations Using GPT-4[22] and Nep Kanun, a RAG-based legal retrieval system[23], stand out as the only dedicated studies focused on Nepali legal information retrieval. However, these works evaluate a limited set of models and lack comprehensive comparative analysis across retrieval approaches. This highlights a significant research gap and underscores the need for further investigation into robust Nepali legal question answering and retrieval systems.

III. METHODOLOGY

A. *Data Collection*

To support the study, we collected two types of data:

a. *Documents:* A total of 10 federal legal documents, as shown in Table I, including the Constitution and nine other widely used federal legal acts, were collected from the Law Commission website[24]. Many of the source files were available only as image-based PDF documents. Therefore, Tesseract OCR[25] was used to convert them into machine-readable text. The extracted text contained OCR-induced artifacts, formatting inconsistencies, and typographical errors, which were subsequently cleaned during preprocessing.

b. *Query Collection*: To evaluate the performance of the retrieval models, a total of 50 query–relevance pairs were constructed, with five queries derived from each of the 10 legal documents (Table I). The queries were collected from social media groups and online legal websites, and were further created, curated, and validated by legal experts to ensure their relevance and clarity. For each query, the corresponding relevant documents were identified to assess the accuracy of different retrieval models.

TABLE I. DOCUMENTS AND QUERIES COUNT

Document	No of queries
Constitution	5
Ecommerce Act	5
Education Act	5
Civil Code	5
Criminal Code	5
Consumer Protection Act	5
Foreign Employment Act	5
Social Security Act	5
Citizenship Act	5
Labor Act	5

B. *Chunking*

To enable accurate retrieval while preserving semantic meaning, documents were systematically chunked into two types: article-level chunks and fixed-size sub-article chunks.

a. *Article-level Chunks:* Articles, being the most meaningful and self-contained units in legal statutes, were initially used as individual chunks. This approach was effective for retrieval models with large context windows.

b. *Fixed-size Sub-Article Chunks:* For models with smaller context sizes, article-level chunks sometimes exceeded the maximum input length, leading to

Identify applicable funding agency here. If none, delete this text box.

information truncation. To address this, each article was further divided into sub-chunks of fewer than 512 tokens, creating fixed-size sub-article chunks suitable for models with limited context windows.

Each chunk was enriched with structured metadata, including the document name, article identifier, and related contextual information, to support retrieval. The chunk structure and metadata format are illustrated in Fig. 1.

```
{
   "text": "संक्षिप्त नाम…..",
   "metadata": {
    "document": "विद्युतिय_व्यापार_(इ-कमर्स)_ऐन",
    "paricched_number": 1,
    "paricched_title": "प्रारम्भिक",
    "dhaara_number": 1,
    "dhaara_title": "संक्षिप्त नाम, प्रारम्भ र बिस्तार",
    "content_type": "dhaara",
    "chunk_index": 0,
    "original_chunk_index": 0
  }
 }
```

Fig. 1. Structure of a chunk before embedding

C. Retrieval Models Selection

For the Nepali legal retrieval analysis, both monolingual and multilingual retrieval models were evaluated. Multilingual models were selected with reference to the (Massive Text Embedding Benchmark)MTEB[1], which provides standardized evaluation across multiple languages and tasks. Using the Hugging Face model hub, we filtered candidate models to include only open-source models with fewer than 1 billion parameters, in accordance with the computational constraints of the available resources (Google Colab). From this pool, seven multilingual models with strong performance were selected for experimentation.

For monolingual retrieval, we evaluated models trained exclusively on the Nepali language. Specifically, we used NepBERTa[26] and its successors based on BERT and RoBERTa architectures, developed by the IRIIS Lab[27]. These models represent the pioneering efforts in Nepali-specific language modeling and serve as strong baselines for evaluating domain-specific retrieval performance. Table II presents all the models used in this study.

TABLE II. SELECTED MODELS

	Model	Type
1.	BAAI/bge-m3	Multi
2.	jina-embeddings-v3	Multi
3.	infloat/multilingual-e5-large-instruct	Multi
4.	Alibaba-NLP/gte-multilingual-base	Multi
5.	Qwen/Qwen3-Embedding-0.6B	Multi
6.	KaLM-embedding-multilingual-mini-instruct-v2	Multi
7.	paraphrase-multilingual-mpnet-base-v2	Multi
8.	NepBERTA	Mono
9.	IRIISNEPAL/RoBERTA	Mono
10.	IRIISNEPAL/BERT	Mono

D. Retrieval Pipeline

For the retrieval pipeline, both small-sized and large-sized chunks were employed. For each chunk type, the chunks and queries were embedded independently using the selected retrieval models. The resulting chunk embeddings were then indexed using FAISS[28], enabling efficient similarity-based search across the corpus. For each query, a nearest-neighbor search based on cosine similarity was performed against the indexed embeddings to retrieve the most relevant document chunks.

The retrieved results were then compared with the annotated relevance labels, and standard retrieval evaluation metrics were computed to assess model performance. The complete retrieval and evaluation workflow, including embedding generation, , query search, and metric computation, is illustrated in the pipeline overview in Fig. II.

IV. EVALUATION

To assess the effectiveness of the retrieval models, we conducted a comprehensive evaluation using standard information retrieval metrics, including Recall, Precision, and Mean Reciprocal Rank (MRR).

a. *Recall*: measures the proportion of relevant documents successfully retrieved by the model out of all relevant documents available. A higher recall indicates that the model is effective in capturing a greater number of relevant items.

b. *Precision:* quantifies the proportion of retrieved documents that are actually relevant. This metric ensures that the model does not return excessive irrelevant results, emphasizing accuracy over quantity.

c. *MRR:* evaluates the rank position of the first relevant document in the retrieved list. This metric is particularly useful in scenarios where presenting a relevant result at the top of the ranking is critical, as it reflects the user experience in practical retrieval settings.

We evaluated 10 retrieval models under two chunking strategies. By employing these metrics, we were able to comprehensively assess not only the accuracy of the models in retrieving relevant documents but also the efficiency of ranking them in an order that maximizes user utility. The results of this evaluation are discussed in the subsequent section.

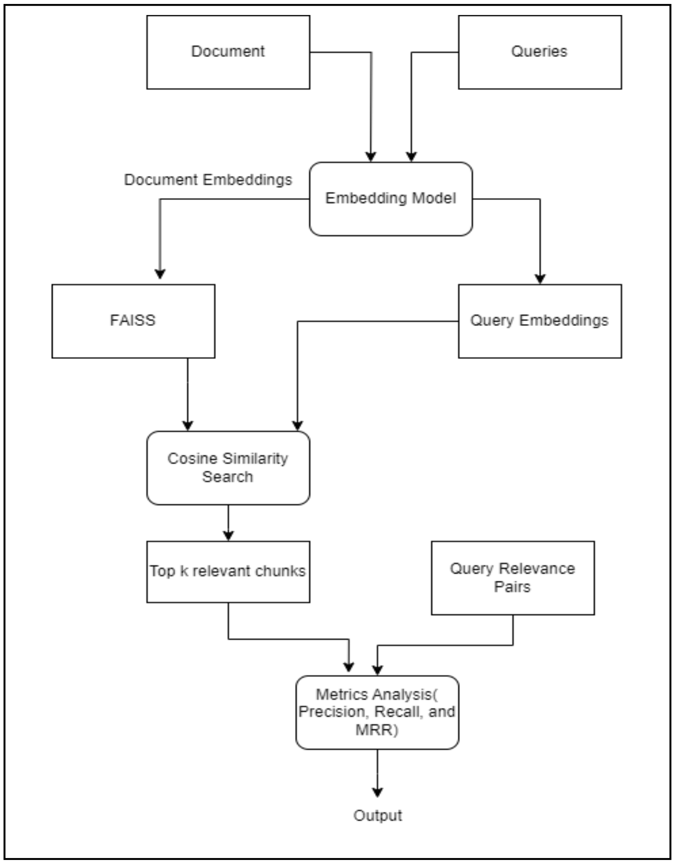

Fig. 2. Retrieval and Analysis Pipeline

V. RESULTS

We evaluated a variety of models, including both monolingual and multilingual approaches, to determine their effectiveness in Nepali legal document retrieval. The results presented below summarize model performance under different conditions using standard evaluation metrics (Recall, Precision, and MRR).

a. *Multilingual models:* Among all evaluated models, BAAI/bge-m3 consistently achieved the highest performance across all metrics with Recall@10 = 0.9233, Precision@1 = 0.7400, and MRR@10 = 0.8300. Jina closely followed in terms of Recall, though its MRR was slightly lower, indicating that while it retrieves relevant documents effectively, the ranking quality is marginally inferior to BGE-m3. Alibaba-GTE and E5-large-instruct achieved moderate performance, trailing behind the top two models (see Appendix A, Tables A.1–A.6).

   Due to the large input context handled by these models, chunking into smaller segments did not produce noticeable differences in the evaluation metrics, suggesting that these models can effectively process long passages without loss of semantic information.

b. *Monolingual Models:* Monolingual models, while somewhat effective, generally underperformed compared to their multilingual counterparts. We evaluated three monolingual models, with the best-performing one being BERT from IRIIS Lab, trained on 27.5GB of Nepali data. Despite this, it achieved Recall@10 = 0.35 and MRR@10 = 0.16, while the other two monolingual models performed significantly worse ((see Appendix A, Tables A.1–A.6)).

   The underperformance of these models can be attributed to the limited size and domain specificity of their training corpora. Since they were trained primarily on general-domain text scraped from news websites rather than legal documents, they were unable to generalize effectively to the legal retrieval task, highlighting the importance of large, domain-specific corpora for monolingual models.

   Additionally, the limitations imposed by large input context sizes can be partially mitigated by using smaller chunks, as demonstrated by BERT's MRR increasing from 0.16 to 0.18 at K=10.

c. *Trend Analysis:* Overall, Recall and MRR increased with higher K values across all models. This trend occurs because, as K increases, the models are allowed to retrieve more documents, which naturally increases the likelihood of including relevant documents (higher Recall) and improves the chances of ranking at least one relevant document higher (higher MRR). In contrast, Precision generally decreased with higher K, as retrieving more documents also increases the proportion of irrelevant ones in the top-K results. The complete results, including all models and chunking strategies, are presented in the Appendix A.

VI. CONCLUSION

In this study, we evaluated a range of multilingual and monolingual models for Nepali legal document retrieval. Our results demonstrate that multilingual models, particularly BAAI/bge-m3, consistently outperform monolingual models across all evaluation metrics and K values. Monolingual models were limited by smaller, general-domain training corpora, which restricted their ability to generalize to legal texts. While chunking smaller input segments slightly improved performance for models with large context limitations, the effect was minimal for high-capacity multilingual models. Overall, our findings indicate that multilingual models are best suited for retrieval tasks in low-resource languages like Nepali, especially when dealing with long and complex legal documents.

VII. FUTURE WORKS

Despite the comprehensive evaluation presented in this study, several directions remain for future work. Expanding the size and diversity of the Nepali legal corpus could further enhance retrieval performance, particularly for monolingual models. Fine-tuning multilingual models on Nepali legal texts may improve their handling of domain-specific terminology and legal nuances. Additionally, integrating retrieval models with generative approaches through retrieval-augmented generation (RAG) could support downstream tasks such as legal question answering and automated summarization. Exploring hybrid models that combine monolingual and multilingual embeddings may also yield performance gains by leveraging both language-specific and cross-lingual features. Finally, user-centric evaluations involving legal practitioners could provide practical insights into system effectiveness in real-world legal workflows.

ACKNOWLEDGMENT

The author gratefully acknowledges Prof. Thakur Prasad Upadhaya for his invaluable guidance, constructive feedback, and continuous encouragement throughout this research. The author also thanks the seniors at Fuse machines for their technical support, insightful discussions, and practical guidance, which significantly contributed to the completion of this work.

REFERENCES

[1] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “MTEB: Massive Text Embedding Benchmark,” in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2014–2037. doi: 10.18653/v1/2023.eacl-main.148.

[2] “Information retrieval.” [Online]. Available: https://en.wikipedia.org/wiki/Information_retrieval

[3] M. Sanderson and W. B. Croft, “The History of Information Retrieval Research,” *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444–1451, May 2012, doi: 10.1109/JPROC.2012.2189916.

[4] M. V. Koroteev, “BERT: A Review of Applications in Natural Language Processing and Understanding,” 2021, *arXiv*. doi: 10.48550/ARXIV.2103.11943.

[5] Z. Chi *et al.*, “XLM-E: Cross-lingual Language Model Pre-training via ELECTRA,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 6170–6182. doi: 10.18653/v1/2022.acl-long.427.

[6] Z. Li, “A Classification Retrieval Approach for English Legal Texts,” in *2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Changsha, China: IEEE, Jan. 2019, pp. 220–223. doi: 10.1109/ICITBS.2019.00059.

[7] N. Zhang, Y.-F. Pu, and P. Wang, “An Ontology-based Approach for Chinese Legal Information Retrieval,” in *Proceedings of The 5th International Conference on Computer Engineering and Networks — PoS(CENet2015)*, Shanghai, China: Sissa Medialab, Oct. 2015, p. 076. doi: 10.22323/1.259.0076.

[8] B. K. Bal, “Structure of Nepali Grammar,” in *Working Papers, PAN Localization, Madan Puraskar Pustakalaya (or just PAN Localization Working Papers)*, Kathmandu, Nepal: Madan Puraskar Pustakalaya, pp. 332–396.

[9] G. Salton, E. A. Fox, and H. Wu, “Extended Boolean information retrieval,” *Commun. ACM*, vol. 26, no. 11, pp. 1022–1036, Nov. 1983, doi: 10.1145/182.358466.

[10] S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” *Found. Trends® Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[11] T. K. Landauer and S. T. Dumais, “Latent semantic analysis,” *Scholarpedia*, vol. 3, p. 4356, 2008.

[12] W. Xing and A. Ghorbani, “Weighted PageRank algorithm,” in *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, Fredericton, NB, Canada: IEEE, 2004, pp. 305–314. doi: 10.1109/DNSR.2004.1344743.

[13] K. W. Church, “Word2Vec,” *Nat. Lang. Eng.*, vol. 23, no. 1, pp. 155–162, Jan. 2017, doi: 10.1017/S1351324916000334.

[14] O. Khattab and M. Zaharia, “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*, Virtual Event China: ACM, Jul. 2020, pp. 39–48. doi: 10.1145/3397271.3401075.

[15] E. M. Voorhees and D. K. Harman, “The text REtrieval conference (TREC): history and plans for TREC-9,” *SIGIR Forum*, vol. 33, pp. 12–15, 1999.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text,” in *Conference on Empirical Methods in Natural Language Processing*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:11816014

[17] D. F. Campos *et al.*, “MS MARCO: A Human Generated MAchine Reading COmprehension Dataset,” *ArXiv*, vol. abs/1611.09268, 2016, [Online]. Available: https://api.semanticscholar.org/CorpusID:1289517

[18] M. Zhou *et al.*, “CCMusic: An Open and Diverse Database for Chinese Music Information Retrieval Research,” *Trans Int Soc Music Inf Retr*, vol. 8, pp. 22–38, 2025.

[19] A. Lefebvre-Brossard, S. Gazaille, and M. C. Desmarais, “Alloprof: a new French question-answer education dataset and its use in an information retrieval case study,” *ArXiv*, vol. abs/2302.07738, 2023, [Online]. Available: https://api.semanticscholar.org/CorpusID:256868592

[20] “Information Retrieval Through Question Answering on Biomedical Research Papers”, [Online]. Available: https://www.researchgate.net/profile/Shushanta-Pudasaini/publication/375011546_Question_Answering_on_Biomedical_Research_Papers_using_Transfer_Learning_on_BERT-Base_Models/links/6560bd5db1398a779dad41c0/Question-Answering-on-Biomedical-Research-Papers-using-Transfer-Learning-on-BERT-Base-Models.pdf

[21] D. R. Ghimire, S. P. Panday, and A. Shakya, “Information Extraction from a Large Knowledge Graph in the Nepali Language,” *Natl. Coll. Comput. Stud. Res. J.*, vol. 3, no. 1, pp. 33–49, Dec. 2024, doi: 10.3126/nccsrj.v3i1.72336.

[22] “ENHANCED-RETRIEVAL-FOR-QA-SYSTEM-TAILORED-FOR-NEPALI-LEGAL-DOCUMENTS-FOCUSING-ON-PSC-EXAMS-USING-GPT-4-AND-RAG-FRAMEWORK.” [Online]. Available: https://www.scribd.com/document/844123533/Revised2-Unmasked-ENHANCED-RETRIEVAL-FOR-QA-SYSTEM-TAILORED-FOR-NEPALI-LEGAL-DOCUMENTS-FOCUSING-ON-PSC-EXAMS-USING-GPT-4-AND-RAG-FRAMEWORK-pdf

[23] “NepKanun: A RAG-Based Nepali Legal Assistant,” 2025, [Online]. Available: https://openreview.net/forum?id=LuXTBI6GSh

[24] “Nepali Legal Documents.” [Online]. Available: https://lawcommission.gov.np/

[25] R. W. Smith, “An Overview of the Tesseract OCR Engine,” *Ninth Int. Conf. Doc. Anal. Recognit. ICDAR 2007*, vol. 2, pp. 629–633, 2007.

[26] S. Timilsina, M. Gautam, and B. Bhattarai, “NepBERTa: Nepali Language Model Trained in a Large Corpus,” in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Online only: Association for Computational Linguistics, 2022, pp. 273–284. doi: 10.18653/v1/2022.aacl-short.34.

[27] P. Thapa, J. Nyachhyon, M. Sharma, and B. K. Bal, “Development of Pre-Trained Transformer-based Models for the Nepali Language,” in *Proceedings of the First Workshop on Challenges in Processing South Asian*

Languages (CHiPSAL 2025), K. Sarveswaran, A. Vaidya, B. Krishna Bal, S. Shams, and S. Thapa, Eds., Abu Dhabi, UAE: International Committee on Computational Linguistics, Jan. 2025, pp. 9–16. [Online]. Available: https://aclanthology.org/2025.chipsal-1.2/

[28] M. Douze *et al.*, “THE FAISS LIBRARY,” *IEEE Trans. Big Data*, pp. 1–17, 2025, doi: 10.1109/TBDATA.2025.3618474.

APPENDIX A

A.1. ARTICLE LEVEL CHUNKING

TABLE A.1.1 Recall

Model/Recall	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.527	0.607	0.687	0.727	0.757	0.767	0.797	0.797	0.817	0.837
bge	0.667	0.807	0.867	0.907	0.913	0.923	0.923	0.923	0.923	0.923
gte	0.547	0.657	0.697	0.767	0.797	0.817	0.837	0.837	0.857	0.857
qwen	0.297	0.507	0.547	0.637	0.677	0.697	0.727	0.747	0.757	0.757
kalm	0.31	0.487	0.567	0.587	0.597	0.637	0.637	0.657	0.677	0.697
jina	0.627	0.787	0.867	0.907	0.917	0.917	0.917	0.917	0.923	0.923
mpnet	0.42	0.51	0.55	0.62	0.63	0.63	0.67	0.71	0.74	0.74
nepberta	0.07	0.08	0.11	0.12	0.14	0.14	0.16	0.16	0.18	0.2
bert	0.08	0.11	0.13	0.17	0.21	0.27	0.31	0.31	0.35	0.35
robert	0.05	0.09	0.13	0.15	0.15	0.15	0.17	0.17	0.17	0.17

TABLE A.1.2 Precision

Model/Precision	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.58	0.36	0.267	0.21	0.176	0.15	0.134	0.118	0.109	0.1
bge	0.74	0.47	0.34	0.265	0.216	0.183	0.157	0.138	0.122	0.11
gte	0.6	0.38	0.267	0.22	0.184	0.157	0.14	0.122	0.111	0.1
qwen	0.34	0.29	0.213	0.185	0.156	0.133	0.12	0.108	0.098	0.088
kalm	0.34	0.28	0.22	0.17	0.14	0.123	0.106	0.095	0.087	0.08
jina	0.72	0.46	0.34	0.265	0.216	0.18	0.154	0.135	0.122	0.11
mpnet	0.48	0.31	0.22	0.185	0.152	0.127	0.114	0.105	0.098	0.088
nepberta	0.1	0.06	0.053	0.045	0.04	0.033	0.031	0.028	0.027	0.026
bert	0.1	0.07	0.053	0.05	0.052	0.053	0.051	0.045	0.044	0.04
robert	0.06	0.05	0.047	0.04	0.032	0.027	0.026	0.022	0.02	0.018

TABLE A.1.3 MRR

Model/MRR	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.58	0.61	0.637	0.647	0.651	0.651	0.656	0.656	0.661	0.663
bge	0.74	0.8	0.82	0.83	0.83	0.83	0.83	0.83	0.83	0.83
gte	0.6	0.65	0.663	0.678	0.682	0.686	0.691	0.691	0.694	0.694
qwen	0.34	0.43	0.45	0.47	0.478	0.481	0.487	0.49	0.492	0.492
kalm	0.34	0.44	0.46	0.465	0.465	0.472	0.472	0.474	0.476	0.478
jina	0.72	0.78	0.807	0.817	0.817	0.817	0.817	0.817	0.817	0.817
mpnet	0.48	0.5	0.513	0.533	0.537	0.537	0.543	0.548	0.552	0.552
nepberta	0.1	0.1	0.113	0.113	0.117	0.117	0.12	0.12	0.122	0.124
bert	0.1	0.12	0.127	0.137	0.145	0.155	0.16	0.16	0.165	0.165
robert	0.06	0.08	0.093	0.098	0.098	0.098	0.101	0.101	0.101	0.101

A.2. FIXED SIZE CHUNKING

TABLE A.2.1 Recall

Model/Recall	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.504	0.665	0.695	0.715	0.715	0.745	0.775	0.781	0.821	0.841
bge	0.614	0.751	0.775	0.865	0.869	0.869	0.889	0.889	0.889	0.899
gte	0.514	0.621	0.675	0.705	0.765	0.805	0.815	0.815	0.825	0.825
qwen	0.294	0.451	0.535	0.605	0.665	0.705	0.705	0.725	0.725	0.725
kalm	0.28	0.44	0.541	0.571	0.581	0.601	0.621	0.621	0.641	0.641
jina	0.584	0.738	0.815	0.875	0.881	0.901	0.901	0.901	0.901	0.905
mpnet	0.417	0.507	0.567	0.577	0.607	0.633	0.673	0.673	0.703	0.723
nepberta	0.07	0.08	0.11	0.12	0.147	0.147	0.147	0.147	0.167	0.187
bert	0.09	0.14	0.167	0.167	0.197	0.247	0.287	0.287	0.297	0.357
roberta	0.07	0.07	0.12	0.16	0.21	0.21	0.24	0.24	0.24	0.24

TABLE A.2.2 Precision

Model/Precision	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.56	0.4	0.28	0.22	0.176	0.153	0.137	0.122	0.116	0.106
bge	0.7	0.46	0.327	0.27	0.22	0.183	0.16	0.14	0.124	0.114
gte	0.58	0.38	0.28	0.22	0.192	0.167	0.146	0.128	0.116	0.104
qwen	0.34	0.27	0.22	0.185	0.16	0.143	0.123	0.112	0.1	0.09
kalm	0.32	0.25	0.213	0.17	0.14	0.12	0.109	0.095	0.087	0.078
jina	0.68	0.46	0.34	0.27	0.22	0.19	0.163	0.142	0.127	0.116
mpnet	0.5	0.32	0.24	0.185	0.156	0.137	0.123	0.108	0.1	0.092
nepberta	0.1	0.06	0.053	0.045	0.044	0.037	0.031	0.028	0.027	0.026
bert	0.12	0.09	0.073	0.055	0.052	0.053	0.051	0.045	0.042	0.044
roberta	0.08	0.04	0.047	0.045	0.048	0.04	0.04	0.035	0.031	0.028

TABLE A.2.3 MRR

Model/Precision	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
e5-large	0.56	0.63	0.643	0.648	0.648	0.652	0.657	0.657	0.664	0.666
bge	0.7	0.76	0.767	0.787	0.787	0.787	0.79	0.79	0.79	0.79
gte	0.58	0.63	0.65	0.655	0.663	0.67	0.67	0.67	0.672	0.672
qwen	0.34	0.41	0.443	0.458	0.47	0.48	0.48	0.483	0.483	0.483
kalm	0.32	0.4	0.44	0.445	0.449	0.452	0.455	0.455	0.457	0.457
jina	0.68	0.74	0.767	0.782	0.782	0.782	0.782	0.782	0.782	0.782
mpnet	0.5	0.53	0.543	0.543	0.551	0.555	0.56	0.56	0.565	0.567
nepberta	0.1	0.1	0.113	0.113	0.121	0.121	0.121	0.121	0.124	0.126
bert	0.12	0.15	0.163	0.163	0.167	0.177	0.183	0.183	0.183	0.189
roberta	0.08	0.08	0.1	0.11	0.118	0.118	0.124	0.124	0.124	0.124