Benchmarking and Building Zero-Shot Hindi Retrieval Model with Hindi-BEIR and NLLB-E5

Arkadeep Acharya[1] *, Rudra Murthy[2], Vishwajeet Kumar[2], Jaydeep Sen[2],
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna,
[2]IBM Research
arkadeep_2101ai41@iitp.ac.in,{rmurthyv,vishk024,jaydesen}@in.ibm.com

Abstract

Given the large number of Hindi speakers worldwide, there is a pressing need for robust and efficient information retrieval systems for Hindi. Despite ongoing research, comprehensive benchmarks for evaluating retrieval models in Hindi are lacking. To address this gap, we introduce the **Hindi-BEIR** benchmark, comprising 15 datasets across seven distinct tasks. We evaluate state-of-the-art multilingual retrieval models on the Hindi-BEIR benchmark, identifying task and domain-specific challenges that impact Hindi retrieval performance. Building on the insights from these results, we introduce **NLLB-E5**, a novel multilingual retrieval model that leverages a zero-shot approach to support Hindi without the need for Hindi training data. We believe our contributions, which include the release of the **Hindi-BEIR** benchmark and the **NLLB-E5** model, will prove to be a valuable resource for researchers and promote advancements in multilingual retrieval models. The datasets from Hindi-BEIR are publicly available at **Hindi-BEIR**.The training and evaluation code for the NLLB-E5 model can be found on GitHub: **NLLB-E5 Repository**.

1 Introduction

Information retrieval (IR) models are indispensable in our digital age, enabling swift and accurate extraction of relevant data from vast amounts of information. These models enable quick and accurate retrieval of relevant data, thereby facilitating informed decision-making across various downstream applications and domains. In this modern era of LLMs, retrievers have become even more relevant and almost necessary in curbing hallucinations from Large Language Model (LLM) generations by means of much popular retrieval augmented generation (RAG) (Lewis et al., 2021) methods. However, the focus of information retrieval (IR) research has predominantly centered on the English language, which can be attributed to the availability of comprehensive evaluation benchmarks, such as the BEIR (Thakur et al., 2021). Although, more recently, there have been efforts to build multi-lingual retrievers and associated evaluation benchmarks (Zhang et al., 2023, 2021) on non-English languages, multi-lingual expansion of robust benchmarks is still a work in progress, with many languages like Hindi lacking a BEIR like a comprehensive benchmark.

Hindi is one of the official languages of India[1], a language of immense importance and global reach, with over half a billion speakers worldwide and ranking as the third most widely spoken language[2]. However, from NLP research point of view, Hindi can be categorized as a low-resource language, lacking comprehensive resource and benchmarks to advance scientific research and this includes lack of a good retrieval benchmark too. Though there has been some efforts in recent time for the development of Indic language dataset for various NLP task like Question answering (Sabane et al. (2024), Doddapaneni et al. (2023a)) and Summarization (Datta et al. (2023), Ghosh et al. (2024a), Ghosh et al. (2024b)), the evolution of Hindi or, in general, Indic language retrieval datasets has been rather ad-hoc, restricted to a small subset of domains, languages and tasks. Although a recent effort attempts to incorporate multiple Indic languages (Haq et al., 2023) for a retrieval benchmark, the absence of diversity in domains and tasks remains a major limitation, affecting the robust evaluation of the current state-of-the-art retriever models in practical applications, which often involve varying tasks and domains.

In this work, we posit that creating a BEIR-like diversified and robust benchmark is the necessary

*Work done as an intern at IBM Research

[1]https://en.wikipedia.org/wiki/Languages_of_India

[2]https://www.thecollector.com/what-are-the-most-spoken-languages-in-the-world/

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4328–4348

first step to track and advance any retrieval research effectively in a new language. Such a benchmark will be more useful than ad-hoc datasets with multiple languages. Thus, we choose the most widely spoken language among Indic languages, Hindi, as our pivot language to pioneer a large-scale diversified retrieval benchmark **Hindi-BEIR**, spanning 15 diverse datasets across 7 tasks. **Hindi-BEIR** aims for two key objectives:(1) Establish a standardized retrieval benchmark to assess, compare, and advance the state-of-the-art retriever models and (2) Provide workable insights into the potential research directions on building retrieval models for Hindi. With multiple tasks and domains in **Hindi-BEIR**, we posit that **Hindi-BEIR** will provide a more realistic and accurate assessment of retriever models for zero-shot usage in unseen domains and applications.

On the choice of Hindi language it is important to note here, besides being a language of high usage, a retrieval benchmark in Hindi offers some unique challenges too such as:

• **Script Difference:** Hindi uses the Devanagari script, which is fundamentally different from the Latin script used in English. This affects character encoding, text normalization, and processing. Existing tokenizers trained on English data may not handle Hindi text well, thus needing to evaluate various tokenization strategies too.

•**Grammatical Structure:** Hindi predominantly uses Subject-Object-Verb (SOV) word-order as opposed to the Subject-Verb-Object (SVO) word-order in English. Hindi words often include more inflections and agglutinations, affecting word tokenization and testing the robustness of retrieval models.

• **Ambiguity:** In Hindi, some proper nouns can also function as common nouns. For instance, the name *Lata*, a common female name, can also refer to *creeper*, a common noun. A query like *lata ko kaise saaph karen* (how to clean a creeper) can easily mislead lexicon-based retrieval systems and test a model's ability for word sense disambiguation.

The introduction of **Hindi-BEIR** benchmark lets us assess and compare the existing multi-lingual retriever models more accurately for their zero-shot performance in Hindi across diverse domains and tasks covered in the benchmark. Based on this performance comparison, we notice that a key limitation of existing models is the need for a substantial amount of language-specific training data, which is often not available for low-resource languages such as Hindi. To address this problem, we propose **NLLB-E5**, a multilingual retrieval model, inspired by recent works like Langbridge (Yoon et al., 2024) and NLLB-LLM2Vec (Schmidt et al., 2024), which distills information from a monolingual retrieval model into a multilingual retrieval model in a zero shot setup via a multi-lingual encoder. The key strength of **NLLB-E5** lies in its ability to perform distillation without reliance on multilingual training data, thereby eliminating the necessity for supervised multilingual datasets.

We summarize our contributions as follows:

- We introduce **Hindi-BEIR**, the first comprehensive retrieval benchmark in the Hindi language, encompassing 15 diverse datasets from 7 distinct tasks. This was accomplished through the harmonization of high-quality quality verified translated data from BEIR, the conversion of existing Hindi datasets for retrieval purposes, and the introduction of strategically generated synthetic data.

- We develop **NLLB-E5**, which combines a multilingual encoder with a monolingual retrieval model through knowledge distillation to develop a multilingual retrieval model without the need for any multilingual training data.

- We evaluate **NLLB-E5** on Hindi-BEIR and compare against existing multilingual models, which have been trained on multilingual data. We empirically demonstrate that **NLLB-E5** achieves strong performance on Hindi-BEIR, outperforming existing multilingual models without the need for multilingual training data.

2 Related Works

One of the known sources of testing retriever performance on Indic languages was to test on the Indic language subset contained in multi-lingual IR datasets such as MIRACL (Zhang et al., 2022), Mr.TyDi (Zhang et al., 2021) having instances from Indic languages such as Hindi, Bengali, and Telugu. However, two major drawbacks here were: (1) all of these subsets were specific to a domain, hence couldn't really be used as a robust benchmark, and (2) the Indic language-specific subsets of instances didn't contain enough data points to facilitate building neural retrieval models for Indic languages either.

More recently, a relatively large-scale retrieval benchmark MS MARCO (Nguyen et al., 2016) got translated into multiple languages to produce mMARCO (Bonifacio et al., 2022), including Hindi from the set of Indic languages. mMARCO was further extended to more Indic languages by IndicIRSuite (Haq et al., 2023), which is the most recent retrieval benchmark in Indic languages. However, the common problem for both mMARCO and IndicMarco is that these benchmarks, too, are tied to a single type of retrieval task and web domain as was originally in MS MARCO and, therefore, do not provide a robust benchmark.

In terms of Language specific benchmarks, Snegirev et al. (2024) introduce the ruMTEB Benchmark, an embedding model evaluation benchmark in Russian with Retrieval being a subtask, Valentini et al. (2024) introduce the Spanish IR dataset while Wojtasik et al. (2024) release the IR Benchmark in the Polish Language. In Indian languages such as Hindi, the Forum for Information Retrieval Evaluation (FIRE)[3] has released numerous sub-tasks and datasets over the years to facilitate the benchmarking and development of information retrieval models. However, these datasets have predominantly been sourced from various news outlets and are not freely or openly accessible to the public.

To the best of our knowledge, **Hindi-BEIR** presents the first comprehensive IR benchmark, which spans diverse domains and tasks and therefore, provides the first BEIR (Thakur et al., 2021) equivalent comprehensive retrieval benchmark in Hindi.

For multilingual retrieval model development, the most promising works are from Chen et al. (2024) and Wang et al. (2024c), but both these models are heavily dependent on multilingual data and do not scale for languages and domains where data is not readily available in the target language.

Yoon et al. (2024) proposed an approach to equip existing LLMs with multilingual capability without any multilingual supervised data. They vertically stack a multilingual encoder and an LLM, adding a projection/alignment layer between them. The input first passes through the multilingual encoder, the representation from the multilingual encoder passes through the alignment layer and then the output from the alignment layer is passed as input to the LLM. They use English-labeled data to fine-tune the overall model and update only the embedding layers and the alignment layer. The model showed better multilingual capabilities.

[3] https://fire.irsi.org.in/fire/2024/home

Schmidt et al. (2024) extended the LLM2vec model (BehnamGhader et al., 2024) by vertically stacking a NLLB model (Team et al., 2022) and LLM2vec model. Similar to LangBridge (Yoon et al., 2024), they add a projection/alignment layer between the two and add LoRA parameters on the LLM. Different from LangBridge, they have a teacher model and a student model. They use LLM2vec as the teacher model and the NLLB-LLM2vec as the student model. The input is passed through both the teacher and the student models. They minimize the mean squared error between the final representations obtained from both the teacher and the student model. They train the model in two stages, the first stage being general alignment between the teacher and the student model, and, the second stage being task alignment. They show that the NLLB-LLM2vec model is able to obtain superior performance on various natural language understanding tasks.

Our **NLLB-E5** model is inspired by Yoon et al. (2024) who propose LangBridge, an approach which enables multilingual reasoning without multilingual supervised training and Schmidt et al. (2024) who opt for a distillation based approach and introduce NLLB-LLM2Vec capable of producing robust multilingual embeddings extending from NLLB encoder.

3 Hindi-BEIR Retrieval Benchmark

Dataset Name	Tasks	#Corpus	# Queries	Average # words	
				Corpus	Query
ArguAna	Argument Retrieval	7763	1194	159.20	178.20
FiQA-2018	Question-Answering	48178	5924	118.02	15.23
TREC-COVID	Bio-Medical IR	76492	49	159.77	14.39
SCIDOCS	Citation-Prediction	22050	850	169.88	12.82
SciFact	Fact-Checking	2849	1099	164.37	19.44
Touché-2020	Argument Retrieval	355273	49	351.47	8.58
NQ	Question Answering	2595865	2952	89.83	9.63
FEVER	Fact Checking	5362876	120075	88.86	9.48
Climate-FEVER	Fact Checking	5362911	1499	88.86	24.40
CC News Retrieval	News Article Retrieval	5005483	49699	272.20	9.30
Sangraha-IR	Question Answering	350000	9744	308.13	30.91
MIRACL	Passage Retrieval	506264	350	66.49	10.42
IndicQARetrieval	Question Answering	261	1544	480.66	10.88
mMARCO	Passage Retrieval	8841823	6980	64.40	6.46
WikiPediaRetreival	Question Answering	13500	1500	77.19	9.54

Table 1: Statistics of the Dataset in the Hindi-BEIR Benchmark showing the tasks which each dataset covers and the number of corpus and query in the evaluation set of each dataset in the Hindi-BEIR Benchmark.

Table 1 summarises the various datasets included in the **Hindi-BEIR** benchmark, along with the number of documents, queries, and domain information.

To ensure that the **Hindi-BEIR** Benchmark is comprehensive, challenging, and accessible to the public for future research and evaluation, we adhered to the following objectives:**1) Diverse Domains and Tasks:** We include datasets from diverse domains like Wikipedia and news articles to niche domains like scientific publications, finance to test the robustness and generalization ability of the retrieval models. Additionally, a good retrieval model should handle documents and queries of varying lengths equally well. We have ensured that the datasets included in the Hindi-BEIR benchmark exhibit a wide range of query and document lengths among the datasets. **2) Difficulty Level:** We ensure that the datasets are challenging and systems relying solely on lexical overlap have a hard time retrieving the correct document. **3) Public Availability:** All datasets curated in **Hindi-BEIR** Benchmark have user-friendly licenses and are publicly available at Hindi-BEIR

We discuss the **Hindi-BEIR** benchmark creation in the following subsections. Please refer to Appendix A.2 for a more detailed analysis of each dataset.

3.1 Translating English Datasets to Hindi:

We translate a subset of the existing English datasets from the BEIR benchmark (Thakur et al., 2021) into Hindi. This approach was necessitated by the need to ensure comprehensive coverage across multiple domains and to maintain a high level of data quality and complexity. We utilized the Indic-Trans2 model [4] (Gala et al., 2023), a multilingual NMT model supporting translations across all 22 scheduled Indic languages (including English). We employ the back-translation technique to retain good translations. Specifically, given an English query/document we translate it to Hindi. This Hindi-translated query/document is then translated back to English. We calculate the Chrf(++) score (Popović, 2017) between the original English query/document and the backtranslated English query/document. We retained only those translations with a Chrf(++) score exceeding a threshold. We empirically set the threshold to 50 after manually verifying the translation quality of texts obtained from different thresholds.

All entries that fell below this threshold were removed. If an entry was part of the corpus and had an associated query, the query was also discarded. Likewise, any corresponding records in the relevancy mapping file were also deleted.

This strategy enables us to leverage the wealth of existing high-quality datasets in English while making them accessible and useful for Hindi language information retrieval tasks. Nine out of the fifteen datasets, which include Arguana, FiQA-2018, TREC-COVID, SCIDOCS, SciFact, Touché-2020, NQ, FEVER, and Climate-FEVER were created by this method.

3.2 Using existing datasets for Retrieval task

We create Hindi CC News Retrieval and Sangraha-IR datasets from existing datasets. We detail the process we followed for each of the datasets as follows:

Hindi CC News Retrieval : This is a cross-lingual dataset derived from the Hindi subset of the Multilingual CC News dataset[5], which contains 7,444,584 data points comprising Hindi news articles and their corresponding titles. The article text, which is in Hindi, serves as the corpus to be retrieved. At the same time, the English news titles act as queries, representing a practical and common real-world scenario that needs the models to have language-agnostic knowledge across both Hindi and English.

We selected only those title-article pairs where the title never appears in the article text, thus reducing the dataset size to 451,803. This ensures no lexical overlap between the title and the article, making the retrieval task challenging as the model needs to rely on cross-lingual cues. From these filtered data points, we randomly selected 49,699 queries to form the test split, which has been incorporated into the Hindi-BEIR benchmark. Additionally, we release training and validation splits consisting of 301,578 and 100,526 query-corpus pairs, respectively. The train, test, and validation splits collectively encompass all data points obtained after the second filtering step.

Sangraha-IR : This dataset builds on top of the Sangraha Dataset (Khan et al., 2024). We handpicked 350,000 entries from the Sangraha-verified subset, focusing on topics that truly represent Indian culture, such as agriculture, Bollywood, cricket, and other similar areas that capture the heart of India. We specifically chose the Sangraha-verified corpus as it consists of scraped data from "human-verified" Websites, OCR-extracted data from high-quality Indic language PDFs, transcribed

[4]Our choice for IndicTrans2 over other translation models has been discussed in A.3

[5]https://huggingface.co/datasets/intfloat/multilingual_cc_news

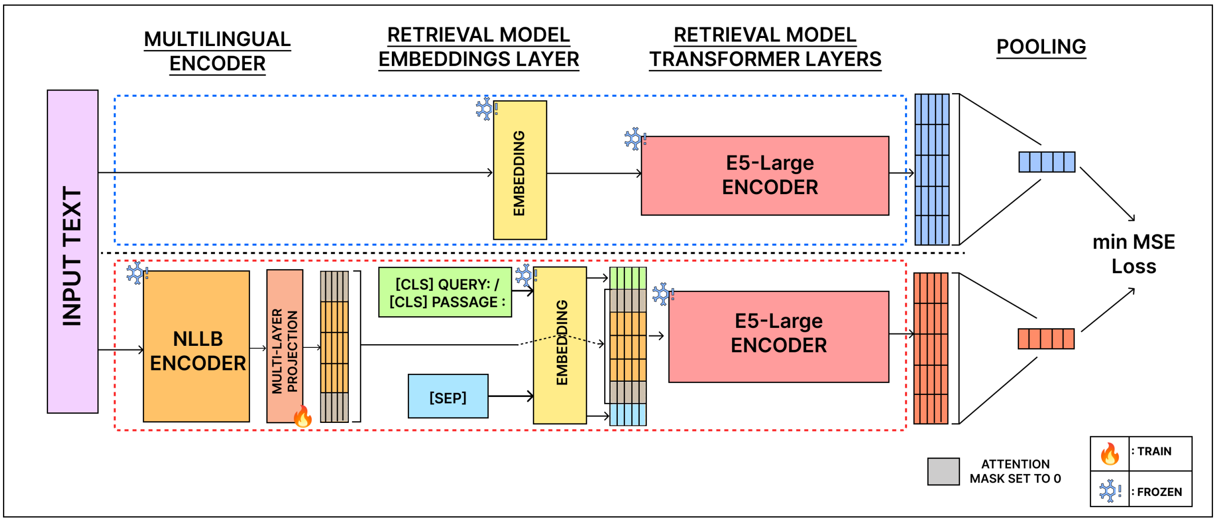


Figure 1: A comprehensive overview of our training methodology. The portion above the dotted lines represents the teacher model (enclosed by a blue dotted box), while the portion below represents our proposed NLLB-E5 student model (enclosed by a red dotted box). We use NLLB as a multilingual encoder and train the Linear projection layer and the Lora adapters by forcing its outputs to match that of the teacher model via the Mean Squared Error (MSE).

data from various Indic language videos, podcasts, movies, courses, etc, and truly captures the essence of the language. To construct a set of relevant queries, we randomly selected 10,000 data points and used the Google Gemini-1.5-flash model to generate corresponding queries [6] for these documents. Queries that did not adhere to the provided instructions or deviated from the desired format were excluded, resulting in a final dataset of 9,744 queries and 350K candidate documents.

3.3 Subset from existing Multilingual Datasets

We also include 5 publicly available datasets, namely MIRACL (dev split)(Zhang et al., 2021), MLDR Chen et al. (2024), mMarco (Bonifacio et al., 2022), IndicQARetrieval (developed by modifying the IndicQA Dataset (Doddapaneni et al., 2023b)) and WikiPediaRetrieval [7].

For translation, we specifically selected a subset of BEIR that focuses on factual and general knowledge content, avoiding topics with strong cultural or language-specific nuances that might be lost in translation. However, to ensure the inclusion of real-world data that captures the linguistic intricacies of Hindi, we carefully incorporate datasets such as Sangraha-IR, CC News Retrieval, MIRACL, IndicQA Retrieval, and Wikipedia Retrieval. These datasets, sourced from manually curated or web-scraped content, effectively preserve the cultural and language-specific nuances essential for robust evaluation.

4 NLLB-E5: multilingual retriever

In this section, we build a strong multilingual retrieval model based on the approach proposed by Yoon et al. (2024) and Schmidt et al. (2024).

Figure 1 shows the architecture of **NLLB-E5** with the different components and data flow between them. We build **NLLB-E5** on the back of two components (1) A multilingual encoder NLLB (Team et al., 2022) and (2) a monolingual retriever E5 (Wang et al., 2024a). We bank on the multilingual encoder to project semantically similar sentences across languages to a shared representation space. We hypothesize that embeddings coming from such a language-agnostic shared representation space can act as a strong prior that can be easily aligned for producing multilingual task-specific embeddings. Moreover, given the multilingual nature of the embedding coming directly from the multilingual encoder, the aforementioned task-specific alignment is language agnostic and, therefore, can be done with task-specific data in a high-resource language, too.

Figure 1 demonstrates how we follow the above intuition in designing **NLLB-E5**. We use NLLB Team et al. (2022) as the strong multilingual encoder that produces robust multilingual embeddings owing to its translation capabilities for over 200 languages. To align these embeddings

[6] Please refer to the Appendix for the prompt used

[7] `https://huggingface.co/datasets/ellamind/wikipedia-2023-11-retrieval-multilingual-queries`

for the retrieval task, we learn a projection layer from the multilingual NLLB encoder to a monolingual retriever E5. The alignment happens by learning the projection layer parameters between NLLB and E5 through distillation from E5 embeddings, which are specifically tuned for retrieval tasks. Since this alignment to the retrieval task is agnostic to a specific language, we use English data and monolingual retriever E5 to do the distillation, eliminating the need for Hindi training data. We chose E5-large as the teacher and student model coupled with NLLB owing to its satisfactory retrieval performance on the BEIR benchmark.

5 Training Methodology

The **NLLB-E5** model is constructed by integrating the monolingual E5 model (Wang et al., 2024a) atop the multilingual encoder from NLLB (Team et al., 2022), connected via a learnable projection layer. The parameters of both the monolingual and multilingual base models remain frozen throughout training. To enable adaptation, we introduce $W \in R^{n \times d}$, a linear projection layer mapping the n-dimensional output space of the multilingual encoder to the d-dimensional input space expected by the E5 model.

We employ knowledge distillation (Hinton et al., 2015) to transfer the capabilities of the monolingual E5 model to the **NLLB-E5** model, facilitating multilingual sentence representation learning. Specifically, the original E5 model serves as the *teacher*, producing high-quality English embeddings, while the **NLLB-E5** model, which includes the learnable Linear projection W, acts as the *student*. By minimizing the discrepancy between the teacher and student embeddings, the student model learns to approximate the E5 model's performance across Hindi and other languages supported by NLLB. The frozen multilingual encoder ensures effective cross-lingual transfer, enabling generalization across all NLLB supported languages. [8]

5.1 Training Process

The training procedure consists of the following steps:

Get the Teacher Model Embeddings: Given an input English sentence S, we first tokenized it using the E5 tokenizer, yielding a sequence of tokens $\{t_1, t_2, \ldots, t_n\}$. These tokens are passed through the teacher E5 model, with frozen parameters, denoted as $\mathcal{M}_T$. The model produces a sequence of token-level representations for the tokens:

$$\{X_1, X_2, \ldots, X_n\} = \mathcal{M}_T(\{t_1, t_2, \ldots, t_n\})$$

where $X_i \in R^d$ represents the d-dimensional embedding of the token t_i produced by the model $\mathcal{M}_T$. To obtain the final sentence-level embedding $\vec{X}_T$, we apply average pooling over the token-level representations, which act as the final embedding for the sentence S obtained from the *teacher model*.

Get Student Model Embeddings: The input sentence is tokenized using the NLLB tokenizer, producing a sequence of tokens $\{t_1, t_2, \ldots, t_m\}$. The frozen multilingual encoder processes this token sequence, generating token-level representations in the multilingual semantic space $\{H_1, H_2, \ldots, H_m\}$. To align with the input structure expected by the E5 model, the token sequence is modified by prepending a token embedding corresponding to either `[CLS] query:` (for queries) or `[CLS] passage:` (for corpus), obtained from the E5 embedding layer (referred from here on as H_{prefix}), along with the representation of `[SEP]`(referred from here on as $H_{postfix}$) which is appended at the end. This is projected through a learnable layer W as $\{H'_1, H'_2, \ldots, H'_m\} = W\{H_{prefix}, H_1, H_2, \ldots, H_m, H_{postfix}\}$ This projected sequence is then passed through the student E5 model, which outputs the token-level representations $\{X'_1, X'_2, \ldots, X'_m\}$ Finally, average pooling is applied over these token-level representations to obtain the final sentence-level embedding $\vec{X}_S$ from the student model.

Minimizing the Loss Function. To align the student model with the teacher, we minimize the Mean Squared Error (MSE) between the teacher and student embeddings, $\vec{X}_T$ and $\vec{X}_S$, respectively, for a batch of sentences.

This optimization encourages the student to generate multilingual embeddings that are aligned with the high-quality English embeddings produced by the teacher.

6 Experiments

This section provides a detailed overview of our experiments on **Hindi-BIER** datasets, including our experimental setup and the baseline models used in our experimentation.

[8]We also experiment with LoRA adapters added on top of the E5 model. Details have been discussed in the Appendix A.5

6.1 Benchmark and Evaluation Metric

We use **Hindi-BEIR** as the target benchmark for all our experiments. We evaluate the models across all 15 datasets in **Hindi-BEIR** and compare their performance at the dataset level too for finer insights.

Following BEIR (Thakur et al., 2021) benchmark standard, we use Normalised Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002), more specifically NDCG@10 as our evaluation metric. NDCG is known to be a more robust metric than simple recall in the IR community because it is purposefully designed to be sensitive towards ranking of the retrieved results.

6.2 Implementation Details of NLLB-E5

To make **NLLB-E5** produce robust, generalizable embeddings across different kind of retrieval tasks and domains, we rely on using diverse data to learn NLLB and E5 alignment during training. We followed a similar approach as taken by generic text-embedding models and used a subset of data from the huge pool of 1B diverse data curated for training sentence transformer models (SIMOULIN, 2021). We randomly selected 100,000 samples from each dataset to create the final training set. The datasets chosen are i) Arxiv ii) Natural Questions (NQ) iii) HotpotQA iv) Stack Exchange - Duplicate questions (titles) v) Stack Exchange - (Title, Body) pairs vi) Stack Exchange - (Title+Body, Answer) pairs vii) Stack Exchange - (Title, Answer) pairs viii) Stack Exchange - Duplicate questions (bodies) ix) Stack Exchange duplicate questions (titles+bodies) x) SQuAD 2.0 xi) S2ORC Citation - pairs (Abstracts) xi) S2ORC (Title, Abstract) xii) SearchQA xiii) PubMed (Title, Abstract) xiv) SPECTER xv) FEVER xvi) Wikipedia Sections xvii) Stack Exchange Math xviii) Stack Overflow Posts xix) Wikipedia

We train the **NLLB-E5** model for 10 epochs, with a learning rate of 2e-4 and a linear scheduler for learning rate adjustment with FP16 precision. We chose the best checkpoint based on the evaluation result on the MsMARCO dev set. We ran all our experiments using Nvidia A100 80GB GPUs. Evaluating the **NLLB-E5** model using 4 A100 80GB GPUs on the **Hindi-BEIR** benchmark took 24 hours.

6.3 Baseline Models

We evaluate our Hindi-BEIR dataset on the following models:

1. The **NLLB-E5 variations:** We develop and evaluate NLLB-E5 using 600M, 1.3B and 3.3B parameter versions of the NLLB model and compare their performance[9].

2. The **BGE-M3** (Chen et al., 2024) model was pre-trained on a large multilingual and cross-lingual unsupervised data and subsequently fine-tuned on high-quality multilingual retrieval datasets using a custom loss function based on the InfoNCE loss function. BGE-M3 supports a context length of 8,192.

3. **Multilingual E5 (mE5)** (Wang et al., 2024c) was developed by continually pre-training the E5 model (Wang et al., 2024b) on a large multilingual corpus using a weakly supervised contrastive pretraining method with InfoNCE contrastive loss van den Oord et al. (2019). It was then fine-tuned on high-quality labelled multilingual datasets for retrieval tasks. mE5 has a context length of 512 tokens.

4. **LASER** (Artetxe and Schwenk, 2019) focuses on universal language agnostic sentence embeddings across 93 languages. LASER uses a language-agnostic BiLSTM encoder architecture trained on parallel corpora from different languages without any retrieval-specific fine-tuning.

5. **LaBSE** (Feng et al., 2022) uses a dual encoder model using BERT to obtain language-agnostic sentence embedding without any retrieval-specific fine-tuning. It has a maximum context length of 256 tokens.

6. **BM-25** (Amati, 2009) is a ranking function in information retrieval based on estimating relevance by combining term frequency and document length normalization.

We extend the MTEB benchmark code repository[10] to include Hindi-BEIR and evaluate the models discussed above.

[9] https://huggingface.co/facebook/nllb-200-distilled-1.3B, https://huggingface.co/facebook/nllb-200-distilled-600M, https://huggingface.co/facebook/nllb-200-3.3B

[10] https://github.com/embeddings-benchmark/mteb

Dataset Name	BGE-M3 (567M)	mE5-Base (110M)	mE5-Large (335M)	LASER	LaBSE (471M)	BM-25	NLLB-E5 (3.3B)
ArguAna	53.81	49.96	54.77	11.27	32.90	43.75	**57.68**
FiQA-2018	25.89	22.38	27.33	1.58	7.23	16.57	**34.41**
TREC-COVID	64.60	62.42	**70.80**	3.95	29.94	52.30	70.12
SCIDOCS	14.24	10.42	11.32	0.59	6.95	11.40	**17.82**
SciFact	52.39	51.50	55.92	5.37	33.42	60.80	**64.34**
Touché-2020	26.68	27.44	26.89	1.06	6.82	**33.59**	25.35
NQ	39.15	44.10	44.10	0.49	9.36	16.79	**53.04**
FEVER	66.91	32.87	39.36	0.19	8.27	40.57	**71.82**
Climate-FEVER	23.71	5.93	8.22	0.28	3.72	14.00	**25.66**
CC News Retrieval	34.40	20.81	**35.00**	0.52	5.63	0.01	31.73
Sangraha-IR	44.85	41.12	**47.91**	3.76	21.76	30.92	42.36
MIRACL	**59.34**	58.11	59.24	0.69	13.76	40.98	53.56
IndicQARetrieval	**69.92**	67.11	67.11	21.28	46.85	74.02	61.94
mMARCO	29.49	29.94	30.92	0.48	6.98	16.53	**33.03**
WikiPediaRetrieval	**87.38**	84.40	86.71	0.03	61.28	82.54	85.65
Average	46.18	40.57	44.37	3.44	19.66	35.65	**48.57**

Table 2: NDCG@10 scores of the existing multilingual model and best-performing NLLB-E5 on Hindi-BEIR datasets. The best-performing model has been highlighted as **bold** while the second-best model has been underlined.

7 Results and Analysis

In this section, we compare the performance of the aforementioned baselines with our proposed **NLLB-E5** model on **Hindi-BEIR** benchmark and highlight our key findings from the results.[11]

Dataset Name	600M (distilled)	1.3B (distilled)	1.3B	3.3B
Arguana	55.23	57.24	56.50	57.68
FiQa-2018	32.83	34.19	33.92	34.41
TREC-COVID	70.07	70.53	70.54	70.12
SCIDOCS	17.53	18.23	17.67	17.82
SciFact	63.93	64.36	64.35	64.34
Touché-2020	25.89	25.21	26.13	25.35
NQ	50.14	53.38	52.77	53.04
FEVER	66.79	71.04	69.02	71.82
Climate-FEVER	25.12	22.51	23.79	25.66
CC News Retrieval	29.11	31.91	30.93	31.73
Sangraha-IR	40.51	43.30	42.84	42.36
MIRACL	50.38	52.96	52.88	53.56
IndicQARetrieval	60.47	62.10	61.69	61.94
mMARCO	31.63	34.03	33.15	33.03
WikiPediaRetrieval	84.89	86.20	86.22	85.65
Average	46.97	48.48	48.16	48.57

Table 3: NDCG@10 scores of NLLB-E5 models using NLLB encoders of different parameter sizes on **Hindi-BEIR** benchmark

NLLB-E5 outperforms multilingual retrievers on Hindi-BEIR: Table 2 shows NLLB$_{1.3B}$-E5 model achieves better average compared to all other baselines on Hindi-BEIR. This is significant because we had state-of-the-art multilingual retriever models such as BGE-M3 and mE5 as our baselines in addition to LASER and LaBSE. It is important to note here that all these multilingual embedding models have seen a huge amount of multilingual data in their training, including Hindi. In contrast, **NLLB-E5** is tuned only on English language data for learning retrieval alignment multilingual embeddings from NLLB encoder.

As we can see in Table 2 **NLLB-E5** performs the best on almost all of BEIR subsets of Hindi-BEIR. More specifically, the **NLLB-E5** model attains the highest scores on the Hindi versions of ArguAna, FiQA-2018, NQ, SCIDOCS and SciFact, FEVER, Climate-FEVER and mMARCO and performs 2nd best on Hindi TREC-COVID dataset. In comparison to its direct competitor, mE5-Large, the NLLB-E5 model demonstrates a substantial edge in fact-checking datasets such as FEVER, Climate-FEVER and SciFact, resulting in a significant boost of 82.47 %, 212 % and 15.06 % on the respective datasets. For non-BIER subsets too, NLLB$_{1.3B}$-E5 easily outperforms LASER and LaBSE and performs competitively with mE5 and BGE-M3.

NLLB-E5 outperforms BM25 on Hindi-BEIR: BM25 is known to be a strong baseline on BEIR (Thakur et al., 2021) benchmark. In fact, as seen in Table 2, BM25 indeed acts as a better baseline than LASER and LaBSE too. However, **NLLB-E5** outperforms BM25 on 13/15 datasets in the Hindi-BEIR benchmark, achieving a gain of $+37\%$ on average. This gain justifies that we

[11]We also evaluated NLLB Encoder trained using distillation approach as used for NLLB-E5. Details has been discussed in A.6

can treat **NLLB-E5** as a robust neural model much better than lexical retrievers such as BM25.

BM25 has very poor performance on CC News Retrieval datasets only. The CC News Retrieval dataset was intentionally designed to test the ability of multilingual dense retrieval models to learn language-agnostic embeddings for retrieval. This dataset presents a scenario where the query is in English, and the corpus is in Hindi, which needs to move beyond token-level matching. This explains the poor performance of BM25 due to the lack of overlap between English and Hindi tokens.

Variation in Performance Across Different Tasks and Domains: One of the important aspects of building Hindi-BEIR was to see how a retriever adapts to different domains and tasks. As we can see, mE5 specifically struggles in fact-checking datasets, while BGE-M3 and NLLB-E5, although not great, still fare better for fact-checking tasks. All the models perform quite poorly for the citation prediction tasks on SCIDOCS. Also, for niche domains such as Finance and Climate, we see a sharp drop for all the models, prompting the need for focused research in these areas.

Effect of Encoder Param Size: Table 3 presents the NDCG@10 scores for the NLLB model's encoder across different parameter sizes. As the table shows, performance generally improves as the parameter size increases, though minor anomalies exist across specific datasets. Notably, the performance gain plateaus beyond the 1.3B parameter model, with only a marginal improvement of 0.09% when comparing the NLLB-E5 model's 1.3B (distilled) encoder to that of the 3.3B parameter version. This possibly indicates that **NLLB-E5** is not lacking in model capacity yet and thus can possibly be improved further by showing more volumes of data from diverse tasks and domains.

8 Conclusion and Future Work

In this work, we introduced **Hindi-BEIR**, a comprehensive benchmark for Hindi language information retrieval. The benchmark consists of 15 datasets, with a corpus size exceeding 27 million and a query size of over 200K, covering a diverse range of 7 tasks. This provides the first-of-its-kind robust benchmark for assessing the performance of retriever models in Hindi. Our experimental results offer insights into the strengths and limitations of current multilingual retrievers on **Hindi-BEIR**, indicating the pressing need for further research in this domain. Additionally, we also proposed the **NLLB-E5** model, a novel architecture for multilingual retriever tuning without needing multilingual training data, capable of handling data from over 200 languages, including Hindi, which outperformed existing multilingual models on Hindi-BEIR.

Future work will focus on expanding the Hindi-BEIR benchmark to include more diversity by curating additional domains such as Law and Medicine. We also plan to extend this benchmark to cover languages beyond Hindi and explore alternative multilingual encoders to optimize representation. We believe our work will have a lasting impact in developing inclusive and scalable information retrieval systems across diverse languages.

9 Limitations

The queries are AI generated etc. While the **Hindi-BEIR** Benchmark and **NLLB-E5** model provides valuable advancements, we acknowledge several limitations. One limitation of the **Hindi-BEIR** Benchmark is that it may not fully capture the breadth of tasks where retrieval models are essential. The current scope, while extensive, lacks coverage in critical domains such as Law and Medicine, which are key areas we plan to include in future expansions. Additionally, although designed primarily for Hindi, the benchmark's extension to other languages is an area that remains unexplored in this work.

The **NLLB-E5** model, while showing promise, exhibits sub-optimal performance on certain English tasks. This limitation indicates a need to explore alternative multilingual encoders to improve language performance. Additionally, we realise that the current model would architecture struggle with long texts due to a context length limitation of 508 tokens, which poses challenges in tasks requiring extended context handling. Further, we would also like to extend and validate this idea for other low-resource languages in the future.

References

Giambattista Amati. 2009. *BM25*, pages 257–260. Springer US, Boston, MA.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions* *of the Association for Computational Linguistics*, 7:597–610.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Debtanu Datta, Shubham Soni, Rajdeep Mukherjee, and Saptarshi Ghosh. 2023. MILDSum: A novel benchmark dataset for multilingual summarization of Indian legal case judgments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5291–5302, Singapore. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023a. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Preprint*, arXiv:2212.05409.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023b. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Preprint*, arXiv:2212.05409.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024a. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024b. HealthAlignSumm : Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560, Miami, Florida, USA. Association for Computational Linguistics.

Saiful Haq, Ashutosh Sharma, and Pushpak Bhattacharyya. 2023. Indicirsuite: Multilingual dataset and neural information models for indian languages. *Preprint*, arXiv:2312.09508.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446.

Mohammed Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh Khapra. 2024. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15831–15879, Bangkok, Thailand. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maithili Sabane, Onkar Litake, and Aman Chadha. 2024. Breaking language barriers: A question answering dataset for hindi and marathi. *Preprint*, arXiv:2308.09862.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual nlu in 200+ languages. *Preprint*, arXiv:2406.12739.

Antoine SIMOULIN. 2021. Train a sentence embedding model with 1 billion training pairs.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *arXiv preprint arXiv:2408.12503*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Preprint*, arXiv:2104.08663.

Francisco Valentini, Viviana Cotik, Damián Furman, Ivan Bercovich, Edgar Altszyler, and Juan Manuel Pérez. 2024. Messirve: A large-scale spanish information retrieval dataset. *Preprint*, arXiv:2409.05994.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59, Copenhagen, Denmark. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024b. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Konrad Wojtasik, Vadim Shishkin, Kacper Wołowiec, Arkadiusz Janz, and Maciej Piasecki. 2024. Beir-pl: Zero shot information retrieval benchmark for the polish language. *Preprint*, arXiv:2305.19840.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. *Preprint*, arXiv:2401.10695.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. Making a miracl: Multilingual information retrieval across a continuum of languages. *Preprint*, arXiv:2210.09984.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

A Appendix

A.1 Dataset License

The datasets will be publicly released with the same license as the parent dataset and will be made available for research purposes.

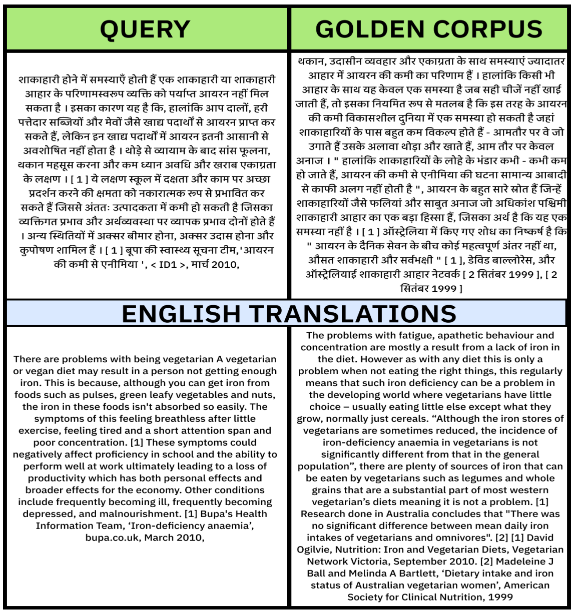
A.2 Dataset Description

A.2.1 ArguAna

1. **Task definition:** Derived from the work by Wachsmuth et al. (2018) , the task is to retrieve the best counterargument, given an argument. Translations of arguments and counter-arguments from online debates constitute the corpus, while the translations of arguments in the original test split, after going through the filtration process based on Chrf++ scores (refer to 3.1), constitute the queries.

2. **Domain :** Misc.

An example of a query with its corresponding golden corpus has been provided in Figure 2



QUERY	GOLDEN CORPUS
शाकाहारी होने में समस्याएँ होती हैं एक शाकाहारी या शाकाहारी आहार के परिणामस्वरूप व्यक्ति को पर्याप्त आयरन नहीं मिल सकता है । इसका कारण यह है कि, हालांकि आप दालों, हरी पत्तेदार सब्जियों और मेवों जैसे खाद्य पदार्थों से आयरन प्राप्त कर सकते हैं, लेकिन इन खाद्य पदार्थों में आयरन इतनी आसानी से अवशोषित नहीं होता है । थोड़े से व्यायाम के बाद सांस फूलना, थकान महसूस करना और कम ध्यान अवधि और खराब एकाग्रता के लक्षण । [1] ये लक्षण स्कूल में दक्षता और काम पर अच्छा प्रदर्शन करने की क्षमता को नकारात्मक रूप से प्रभावित कर सकते हैं जिससे अंततः उत्पादकता में कमी हो सकती है जिसका व्यक्तिगत प्रभाव और अर्थव्यवस्था पर व्यापक प्रभाव दोनों होते हैं । अन्य स्थितियों में अक्सर बीमार होना, अक्सर उदास होना और कुपोषण शामिल हैं । [1] बूपा की स्वास्थ्य सूचना टीम,'आयरन की कमी से एनीमिया ', < ID1 >, मार्च 2010,	थकान, उदासीन व्यवहार और एकाग्रता के साथ समस्याएं ज्यादातर आहार में आयरन की कमी का परिणाम हैं । हालांकि किसी भी आहार के साथ यह केवल एक समस्या है जब सही चीजें नहीं खाई जाती हैं, तो इसका नियमित रूप से मतलब है कि इस तरह के आयरन की कमी विकासशील दुनिया में एक समस्या हो सकती है जहां शाकाहारियों के पास बहुत कम विकल्प होते हैं - आमतौर पर वे जो उगाते हैं उसके अलावा थोड़ा और खाते हैं, आम तौर पर केवल अनाज । " हालांकि शाकाहारियों के लोहे के भंडार कभी - कभी कम हो जाते हैं, आयरन की कमी से एनीमिया की घटना सामान्य आबादी से काफी अलग नहीं होती है ", आयरन के बहुत सारे स्रोत हैं जिन्हें शाकाहारियों जैसे फलियां और साबुत अनाज जो अधिकांश पश्चिमी शाकाहारी आहार का एक बड़ा हिस्सा हैं, जिसका अर्थ है कि यह एक समस्या नहीं है । [1] ऑस्ट्रेलिया में किए गए शोध का निष्कर्ष है कि " आयरन के दैनिक सेवन के बीच कोई महत्वपूर्ण अंतर नहीं था, औसत शाकाहारी और सर्वभक्षी " [1], डेविड बाल्लोरेस, और ऑस्ट्रेलियाई शाकाहारी आहार नेटवर्क [2 सितंबर 1999], [2 सितंबर 1999]
ENGLISH TRANSLATIONS	
There are problems with being vegetarian A vegetarian or vegan diet may result in a person not getting enough iron. This is because, although you can get iron from foods such as pulses, green leafy vegetables and nuts, the iron in these foods isn't absorbed so easily. The symptoms of this feeling breathless after little exercise, feeling tired and a short attention span and poor concentration. [1] These symptoms could negatively affect proficiency in school and the ability to perform well at work ultimately leading to a loss of productivity which has both personal effects and broader effects for the economy. Other conditions include frequently becoming ill, frequently becoming depressed, and malnourishment. [1] Bupa's Health Information Team, 'Iron-deficiency anaemia', bupa.co.uk, March 2010,	The problems with fatigue, apathetic behaviour and concentration are mostly a result from a lack of iron in the diet. However as with any diet this is only a problem when not eating the right things, this regularly means that such iron deficiency can be a problem in the developing world where vegetarians have little choice – usually eating little else except what they grow, normally just cereals. "Although the iron stores of vegetarians are sometimes reduced, the incidence of iron-deficiency anaemia in vegetarians is not significantly different from that in the general population", there are plenty of sources of iron that can be eaten by vegetarians such as legumes and whole grains that are a substantial part of most western vegetarian's diets meaning it is not a problem. [1] Research done in Australia concludes that "There was no significant difference between mean daily iron intakes of vegetarians and omnivores". [2] [1] David Ogilvie, Nutrition: Iron and Vegetarian Diets, Vegetarian Network Victoria, September 2010. [2] Madeleine J Ball and Melinda A Bartlett, 'Dietary intake and iron status of Australian vegetarian women', American Society for Clinical Nutrition, 1999

Figure 2: An example of a query with its corresponding golden corpus from the ArguAna Dataset

Distribution of the number of words in the corpus and queries in the ArguAna dataset has been shown in Figure 4 and Figure 3, respectively.

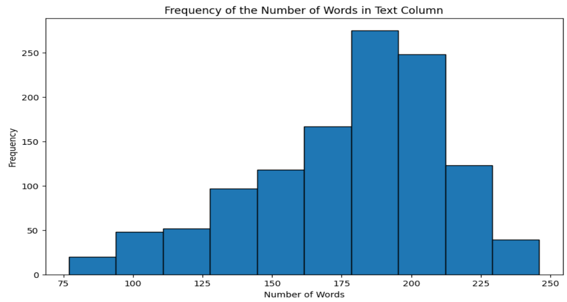


Figure 3: Distribution of the number of words in the queries of ArguAna Dataset

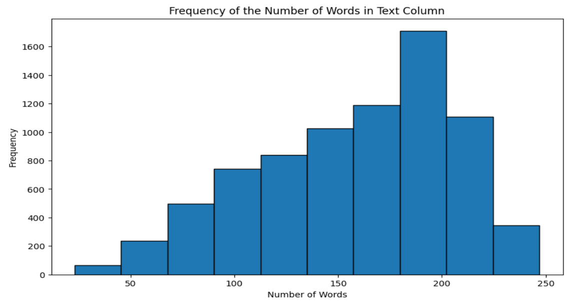


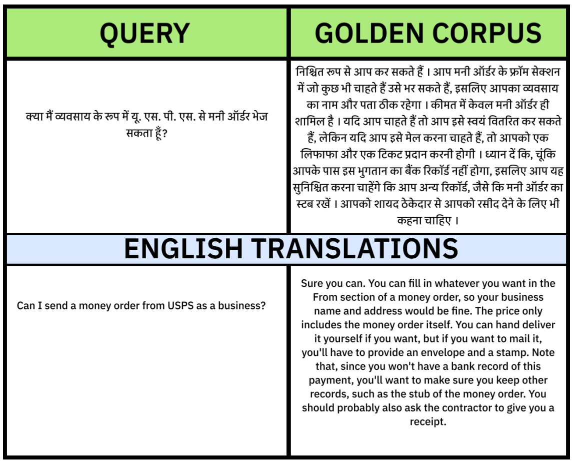
Figure 4: Distribution of Number of Words in the corpus of ArguAna Dataset

A.2.2 FiQA-2018

1. **Task Definition:** It deals with Opinion-Based Question answering. Based on the works of Maia et al. (2018),translation of the financial data extracted by crawling StackExchange posts under the Investment topic from 2009-2017, after passing through filteration processes mentioned in 3.1, acts as the corpus. While translations from the original training split acts as the queries.

2. **Domain :** Finance

An example of query with its corresponding golden corpus from the FiQA-2018 dataset has been provided in Figure 5



QUERY	GOLDEN CORPUS
क्या मैं व्यवसाय के रूप में यू. एस. पी. एस. से मनी ऑर्डर भेज सकता हूँ?	निश्चित रूप से आप कर सकते हैं । आप मनी ऑर्डर के फ्रॉम सेक्शन में जो कुछ भी चाहते हैं उसे भर सकते हैं, इसलिए आपका व्यवसाय का नाम और पता ठीक रहेगा । कीमत में केवल मनी ऑर्डर ही शामिल है । यदि आप चाहते हैं तो आप इसे स्वयं वितरित कर सकते हैं, लेकिन यदि आप इसे मेल करना चाहते हैं, तो आपको एक लिफाफा और एक टिकट प्रदान करनी होगी । ध्यान दें कि, चूंकि आपके पास इस भुगतान का बैंक रिकॉर्ड नहीं होगा, इसलिए आप यह सुनिश्चित करना चाहेंगे कि आप अन्य रिकॉर्ड, जैसे कि मनी ऑर्डर का स्टब रखें । आपको शायद ठेकेदार से आपको रसीद देने के लिए भी कहना चाहिए ।
ENGLISH TRANSLATIONS	
Can I send a money order from USPS as a business?	Sure you can. You can fill in whatever you want in the From section of a money order, so your business name and address would be fine. The price only includes the money order itself. You can hand deliver it yourself if you want, but if you want to mail it, you'll have to provide an envelope and a stamp. Note that, since you won't have a bank record of this payment, you'll want to make sure you keep other records, such as the stub of the money order. You should probably also ask the contractor to give you a receipt.

Figure 5: An example of a query with its corresponding golden corpus from the FiQA-2018 Dataset

Distribution of the number of words in the corpus and queries in FiQA-2018 dataset has been shown in Figure 7 and Figure 6 respectively.

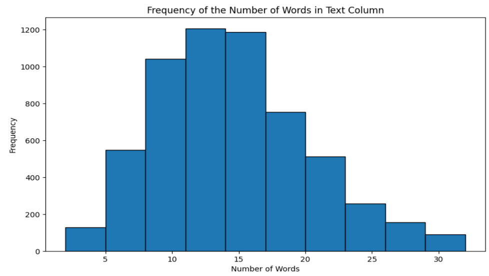


Figure 6: Distribution of the number of words in the queries of FiQA-2018 Dataset

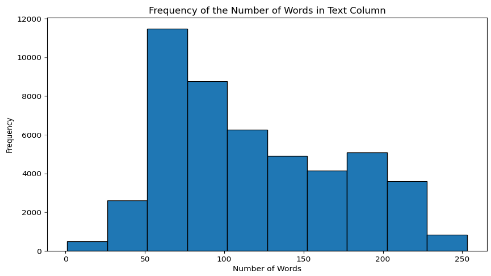


Figure 7: Distribution of Number of Words in corpus of FiQA-2018 Dataset

A.2.3 TREC-COVID

1. **Task Definition:** Voorhees et al. (2021) introduced the original TREC-COVID dataset which is an ad-hoc seach challenge based on CORD-19 dataset containing articles about the COVID-19 Pandemic. The translated and filtered version of the CORD-19 Dataset constitutes the corpus while the final cumulative judgements with query descriptions from the original task are the queries in the TREC-COVID Dataset.

2. **Domain :** Bio-Medical

An example of a query with its corresponding golden corpus from the TREC-COVID dataset has been provided in Figure 8

Distribution of the number of words in the corpus and queries in TREC-COVID dataset has been shown in Figure 10 and Figure 9 respectively.

A.2.4 SCIDOCS

1. **Task Definition:** Inspired by Cohan et al. (2020), in this task, we expect the model to retrieve cited papers for a given scientific paper abstract as input. The corpus contains about 22k translated and filtered scientific paper abstracts and 850 translated paper titles as queries.

2. **Domain :** Scientific

Figure 8: An example of a query with its corresponding golden corpus from the TREC-COVID Dataset

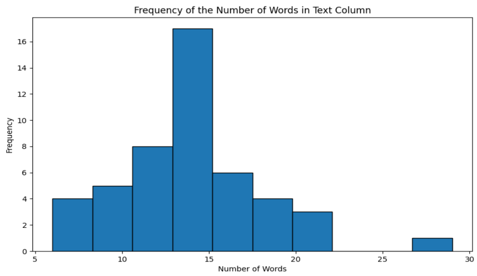


Figure 9: Distribution of the number of words in the queries of TREC-COVID Dataset

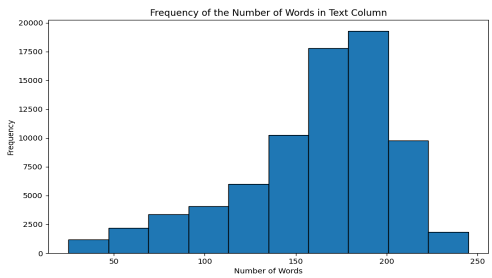
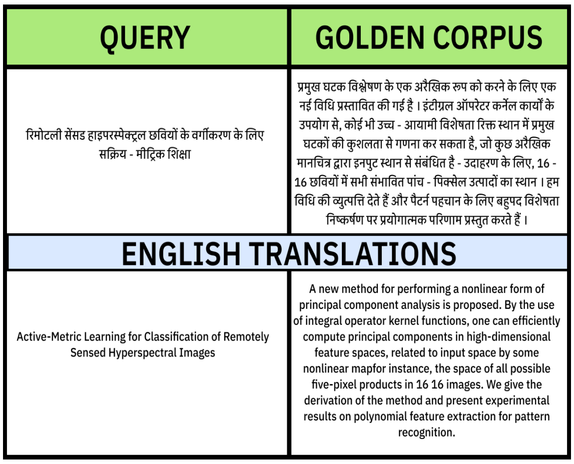


Figure 10: Distribution of Number of Words in the corpus of TREC-COVID Dataset

An example of query with its corresponding golden corpus from the SCIDOCS dataset has been provided in Figure 11



QUERY	GOLDEN CORPUS
रिमोटली सेंसड हाइपरस्पेक्ट्रल छवियों के वर्गीकरण के लिए सक्रिय - मीट्रिक शिक्षा	प्रमुख घटक विश्लेषण के एक अरैखिक रूप को करने के लिए एक नई विधि प्रस्तावित की गई है । इंटीग्रल ऑपरेटर कर्नेल कार्यों के उपयोग से, कोई भी उच्च - आयामी विशेषता रिक्त स्थान में प्रमुख घटकों की कुशलता से गणना कर सकता है, जो कुछ अरैखिक मानचित्र द्वारा इनपुट स्थान से संबंधित है - उदाहरण के लिए, 16 - 16 छवियों में सभी संभावित पांच - पिक्सेल उत्पादों का स्थान । हम विधि की व्युत्पत्ति देते हैं और पैटर्न पहचान के लिए बहुपद विशेषता निष्कर्षण पर प्रयोगात्मक परिणाम प्रस्तुत करते हैं ।
ENGLISH TRANSLATIONS	
Active-Metric Learning for Classification of Remotely Sensed Hyperspectral Images	A new method for performing a nonlinear form of principal component analysis is proposed. By the use of integral operator kernel functions, one can efficiently compute principal components in high-dimensional feature spaces, related to input space by some nonlinear mapfor instance, the space of all possible five-pixel products in 16 16 images. We give the derivation of the method and present experimental results on polynomial feature extraction for pattern recognition.

Figure 11: An example of a query with its corresponding golden corpus from the SCIDOCS Dataset

Distribution of the number of words in the corpus and queries in the SCIDOCS dataset has been shown in Figure 13 and Figure 12, respectively.

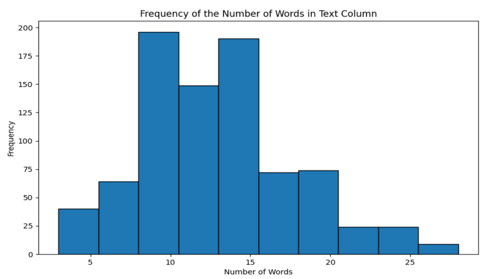


Figure 12: Distribution of the number of words in the queries of SCIDOCS Dataset

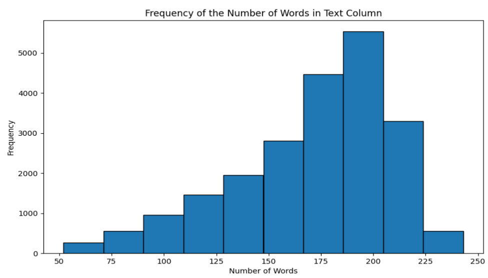
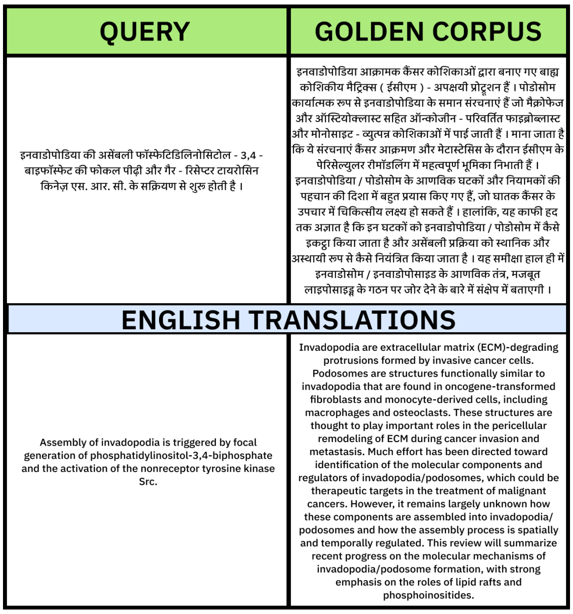


Figure 13: Distribution of Number of Words in the corpus of SCIDOCS Dataset

A.2.5 SciFact

1. **Task Definition:** This task involves the verification of scientific claims given the abstract of scientific articles from recent literature. For this task, the model is expected to retrieve relevant abstracts with a given claim as input.
2. **Domain :** Scientific

An example of query with its corresponding golden corpus from the SciFact dataset has been provided in Figure 14



QUERY	GOLDEN CORPUS
इनवाडोपोडिया की असेंबली फॉस्फेटिडिलिनोसिटोल - 3,4 - बाइफॉस्फेट की फोकल पीढ़ी और गैर - रिसेप्टर टायरोसिन किनेज़ एस. आर. सी. के सक्रियण से शुरू होती है ।	इनवाडोपोडिया आक्रामक कैंसर कोशिकाओं द्वारा बनाए गए बाह्य कोशिकीय मैट्रिक्स (ईसीएम) - अपक्षयी प्रोट्रूशन हैं । पोडोसोम कार्यात्मक रूप से इनवाडोपोडिया के समान संरचनाएं हैं जो मैक्रोफेज और ऑस्टियोक्लास्ट सहित ऑन्कोजीन - परिवर्तित फाइब्रोब्लास्ट और मोनोसाइट - व्युत्पन्न कोशिकाओं में पाई जाती हैं । माना जाता है कि ये संरचनाएं कैंसर आक्रमण और मेटास्टेसिस के दौरान ईसीएम के पेरिसेल्युलर रीमॉडलिंग में महत्वपूर्ण भूमिका निभाती हैं । इनवाडोपोडिया / पोडोसोम के आणविक घटकों और नियामकों की पहचान की दिशा में बहुत प्रयास किए गए हैं, जो घातक कैंसर के उपचार में चिकित्सीय लक्ष्य हो सकते हैं । हालांकि, यह काफी हद तक अज्ञात है कि इन घटकों को इनवाडोपोडिया / पोडोसोम में कैसे इकट्ठा किया जाता है और असेंबली प्रक्रिया को स्थानिक और अस्थायी रूप से कैसे नियंत्रित किया जाता है । यह समीक्षा हाल ही में इनवाडोसोम / इनवाडोपोसाइड के आणविक तंत्र, मजबूत लाइपोसाइड्स के गठन पर जोर देने के बारे में संक्षेप में बताएगी ।
ENGLISH TRANSLATIONS	
Assembly of invadopodia is triggered by focal generation of phosphatidylinositol-3,4-biphosphate and the activation of the nonreceptor tyrosine kinase Src.	Invadopodia are extracellular matrix (ECM)-degrading protrusions formed by invasive cancer cells. Podosomes are structures functionally similar to invadopodia that are found in oncogene-transformed fibroblasts and monocyte-derived cells, including macrophages and osteoclasts. These structures are thought to play important roles in the pericellular remodeling of ECM during cancer invasion and metastasis. Much effort has been directed toward identification of the molecular components and regulators of invadopodia/podosomes, which could be therapeutic targets in the treatment of malignant cancers. However, it remains largely unknown how these components are assembled into invadopodia/ podosomes and how the assembly process is spatially and temporally regulated. This review will summarize recent progress on the molecular mechanisms of invadopodia/podosome formation, with strong emphasis on the roles of lipid rafts and phosphoinositides.

Figure 14: An example of a query with its corresponding golden corpus from the SciFact Dataset

Distribution of the number of words in the corpus and queries in the SciFact dataset has been shown in Figure 16 and Figure 15, respectively.

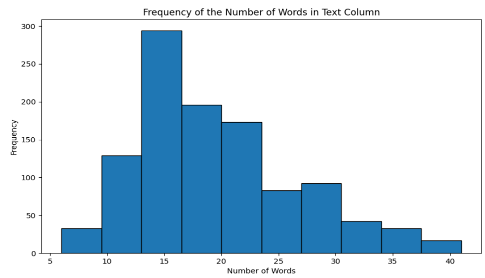


Figure 15: Distribution of the number of words in the queries of SciFact Dataset

A.2.6 Touché-2020

1. **Task Definition:** Similar to ArguAna this task deals with this task deals with the retrieval of conversational arguments. The translated and filtered conclusion forms the title and premise for arguments (Wachsmuth et al.,

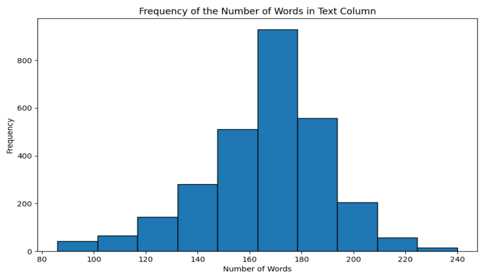


Figure 16: Distribution of Number of Words in the corpus of SciFact Dataset

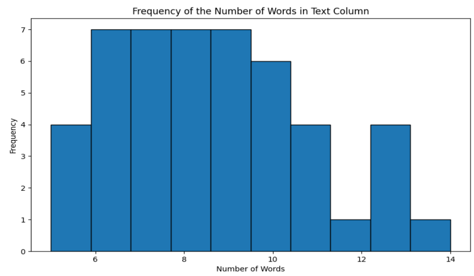


Figure 18: Distribution of the number of words in the queries of Touché-2020 Dataset

2017) constitutes the corpus. The translations of the Touché-2020 task data are the queries.

2. **Domain :** Miscellaneous

An example of query with its corresponding golden corpus from the Touché-2020 dataset has been provided in Figure 17

QUERY	GOLDEN CORPUS
क्या पशुओं का उपयोग वैज्ञानिक या वाणिज्यिक परीक्षण के लिए किया जाना चाहिए?	जानवरों में मनुष्यों पर दवा परीक्षण को अमानवीय या उचित नहीं माना जाता है। लोग दवा परीक्षण पर गुस्सा करते हैं क्योंकि वे डरते हैं कि इससे परीक्षण किए जा रहे व्यक्ति को नुकसान हो सकता है। दवा परीक्षण इस दुनिया में हमारे अधिकांश उपचारों की नींव है। आज हमारे पास कई उपचार दवा परीक्षण पर आधारित हैं। विज्ञान में हम जो कुछ भी जानते हैं वह दवा परीक्षण से आता है। जिन रोगियों पर परीक्षण किया जाता है वे इसे नसों में नहीं करते हैं, यह भविष्य में ठीक होने वाली बीमारियों के इलाज में मदद करने के लिए है
ENGLISH TRANSLATIONS	
Should animals be used for scientific or commercial testing?	Drug testing on humans in animals is not considered inhumane or appropriate. People resent drug testing because they fear it may harm the person being tested. Drug testing is the foundation of most of our treatments in this world. Many of the treatments we have today are based on drug testing. Everything we know in science comes from drug testing. The patients it is tested on do not receive it intravenously, it is to help treat diseases that may be cured in the future.

Figure 17: An example of a query with its corresponding golden corpus from the Touché-2020 Dataset

Distribution of the number of words in the corpus and queries in the Touché-2020 dataset has been shown in Figure 19 and Figure 18 respectively.

A.2.7 NQ

1. **Task Definition:** The task here is to retrieve the correct answer to a given question from a broad corpus of candidate answers. The NQ dataset in the Hindi-BEIR Benchmark is the translated and filtered version of the NQ dataset released by Thakur et al. (2021), which contains Google search queries and documents with paragraphs and answer spans within Wikipedia articles as the corpus.

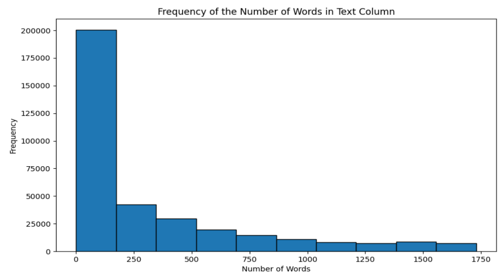


Figure 19: Distribution of Number of Words in the corpus of Touché-2020 Dataset

2. **Domain :** WikiPedia

An example of a query with its corresponding golden corpus from the NQ dataset has been provided in Figure 20

QUERY	GOLDEN CORPUS
एफिल मीनार के प्रत्येक स्तर पर क्या है	टावर में आगंतुकों के लिए तीन स्तर हैं, जिसमें पहले और दूसरे स्तर पर रेस्तरां हैं । शीर्ष स्तर का ऊपरी मंच जमीन से 276 मीटर (906 फीट) ऊपर है - यूरोपीय संघ में जनता के लिए सुलभ सबसे ऊंचा अवलोकन डेक । पहले और दूसरे स्तर तक सीढ़ियों या लिफ्ट (लिफ्ट) से चढ़ने के लिए टिकट खरीदे जा सकते हैं । जमीन के स्तर से पहले स्तर तक चढ़ाई 300 से अधिक सीढ़ियां हैं, जैसा कि पहले स्तर से दूसरे स्तर तक चढ़ाई है । हालांकि शीर्ष स्तर तक एक सीढ़ी है, यह आमतौर पर केवल लिफ्ट द्वारा पहुँचा जा सकता है ।
ENGLISH TRANSLATIONS	
what's on each level of the eiffel tower	The tower has three levels for visitors, with restaurants on the first and second levels. The top level's upper platform is 276 m (906 ft) above the ground – the highest observation deck accessible to the public in the European Union. Tickets can be purchased to ascend by stairs or lift (elevator) to the first and second levels. The climb from ground level to the first level is over 300 steps, as is the climb from the first level to the second. Although there is a staircase to the top level, it is usually accessible only by lift.

Figure 20: An example of a query with its corresponding golden corpus from the NQ Dataset

Distribution of the number of words in the corpus and queries in the NQ dataset has been shown in Figure 22 and Figure 21, respectively.

A.2.8 FEVER

1. **Task Definition:** Similar to SciFact, here, the task is to retrieve relevant documents that

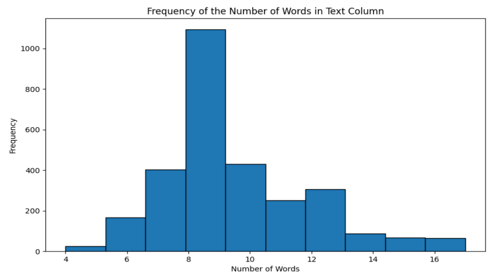


Figure 21: Distribution of the number of words in the queries of NQ Dataset

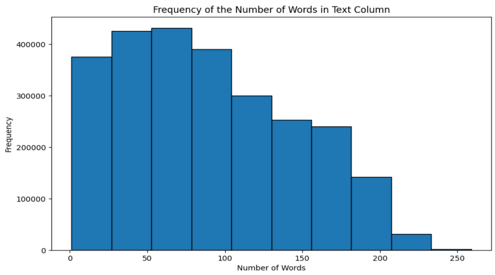


Figure 22: Distribution of Number of Words in the corpus of NQ Dataset

claim a given fact (query). We translate and filter the test split of the FEVER dataset as proposed by Thakur et al. (2021) and include it in the Hindi-BEIR Benchmark.

2. **Domain :** Wikipedia

An example of a query with its corresponding golden corpus from the FEVER dataset has been provided in Figure 23

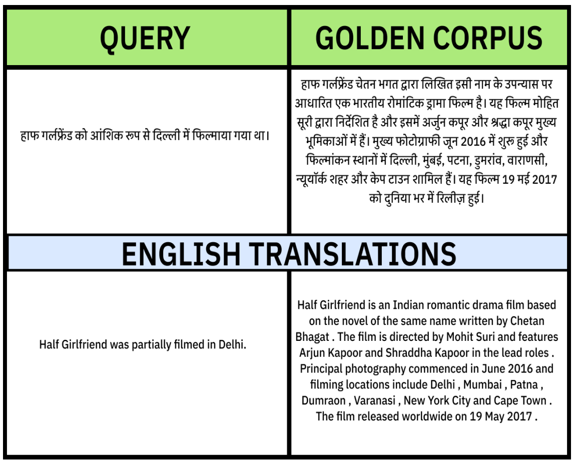


Figure 23: An example of a query with its corresponding golden corpus from the FEVER Dataset

Distribution of the number of words in the corpus and queries in the FEVER dataset has been shown in Figure 25 and Figure 24, respectively.

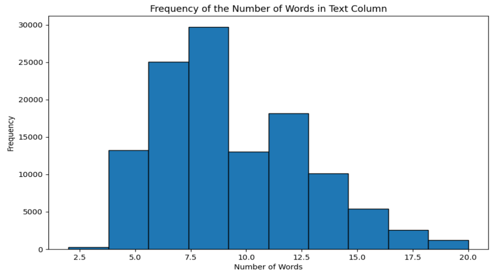


Figure 24: Distribution of the number of words in the queries of FEVER Dataset

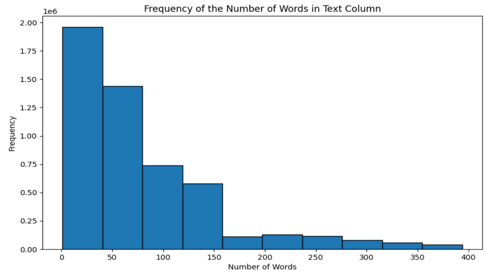


Figure 25: Distribution of Number of Words in the corpus of FEVER Dataset

A.2.9 Climate-FEVER

1. **Task Definition:** Similar to FEVER, Climate-FEVER is a dataset for the verification of real-world climate claims. We translate and filter the test split of the Climate-FEVER dataset as proposed by Thakur et al. (2021) and include it in the Hindi-BEIR Benchmark.

2. **Domain :** Wikipedia

An example of query with its corresponding golden corpus from the Climate-FEVER dataset has been provided in Figure 26

Distribution of the number of words in the corpus and queries in the Climate-FEVER dataset has been shown in Figure 28 and Figure 27, respectively.

A.2.10 CC News Retrieval

1. **Task Definition:** CC News Retrieval Introduces the task to retrieving relevant news articles given a news title. Subsection 3.2 talks in detail about the data creation process.

2. **Domain :** News

QUERY	GOLDEN CORPUS
अंटार्कटिका समुद्री बर्फ का अध्ययन करने वाले वैज्ञानिकों ने चेतावनी दी है कि संचय में वृद्धि अगले हिम युग को जन्म दे सकती है।	एक हिम युग पृथ्वी की सतह और वायुमंडल के तापमान में दीर्घकालिक कमी की अवधि है, जिसके परिणामस्वरूप महाद्वीपीय और ध्रुवीय बर्फ की चादरों और अल्पाइन ग्लेशियरों की उपस्थिति या विस्तार होता है। एक दीर्घकालिक हिम युग के भीतर, ठंडी जलवायु की अलग-अलग दालों को "हिमनदीय अवधि" कहा जाता है (या वैकल्पिक रूप से "हिमनदों" या "हिमनदों" या बोलचाल की भाषा में "हिम युग" के रूप में), और रुक-रुक कर गर्म अवधि को "इंटरग्लेसियल" कहा जाता है। हिमविज्ञान की शब्दावली में, हिम युग उत्तरी और दक्षिणी गोलार्धों दोनों में व्यापक बर्फ की चादरों की उपस्थिति का संकेत देता है। इस परिभाषा के अनुसार, हम एक इंटरग्लेसियल अवधि में हैं-होलोसीन-बर्फ युग की। 26 लाख साल पहले प्लीस्टोसिन युग की शुरुआत में, क्योंकि ग्रीनलैंड, आर्कटिक और अंटार्कटिक बर्फ की चादरें अभी भी मौजूद हैं।
ENGLISH TRANSLATIONS	
Scientists studying Antarctica sea ice warn a rise in accumulation could spark the next ice age.	An ice age is a period of long-term reduction in the temperature of Earth 's surface and atmosphere , resulting in the presence or expansion of continental and polar ice sheets and alpine glaciers . Within a long-term ice age , individual pulses of cold climate are termed glacial periods " (or alternatively glacials " or glaciations " or colloquially as ice age ") , and intermittent warm periods are called `` interglacials " . In the terminology of glaciology , ice age implies the presence of extensive ice sheets in both northern and southern hemispheres . By this definition , we are in an interglacial period -- the Holocene -- of the ice age . The ice age began 2.6 million years ago at the start of the Pleistocene epoch , because the Greenland , Arctic , and Antarctic ice sheets still exist .

Figure 26: An example of a query with its corresponding golden corpus from the Climate-FEVER Dataset

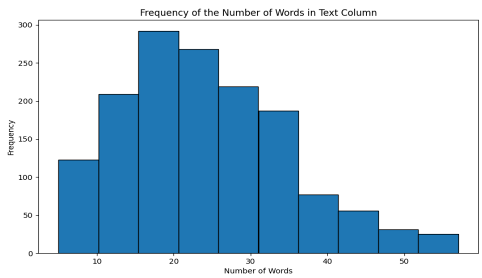


Figure 27: Distribution of the number of words in the queries of Climate-FEVER Dataset

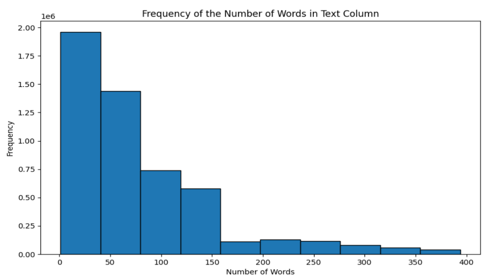


Figure 28: Distribution of Number of Words in corpus of Climate-FEVER Dataset

An example of a query with its corresponding golden corpus from the CC News Retrieval dataset has been provided in Figure 29

Distribution of the number of words in the corpus and queries in CC News Retrieval dataset has been shown in Figure 31 and Figure 30 respectively.

QUERY	GOLDEN CORPUS
Home Remedies For Dry Hair	यूटिलिटी डेस्क।सर्दी में कई लोगों को ड्राय हेयर की प्रॉब्लम होने लगती है। ऐसे में बाल रूखे और बेजान दिखने लगते हैं। लेकिन इस प्रॉब्लम को हम कुछ घरेलू उपायों के जरिए दूर कर सकते हैं। यहां इस वीडियो के जरिए हम आपको बता रहे हैं ऐसे ही कुछ घरेलू उपाय जिन्हें सप्ताह में 3 या 4 बार अपनाकर ड्राय हेयर की प्रॉब्लम को कम किया जा सकता है और बालों को शाइनी बनाया जा सकता है। ऑलिव आयल 3 से 4 चम्मच ऑलिव ऑयल को गर्म कर लें। गुनगुना होने पर इससे बाल पर मसाज करें। इसके बाद तौलिये को हल्का गर्म करके बालों पर बांध लें। एक घंटे बाद नहा लें। नारियल पानी 3 चम्मच नारियल पानी में एक चम्मच नींबू पानी मिला लें। इसे बालों पर लगाएं। 5 मिनट तक भाप दें। फिर एक घंटे बाद नहा लें। अंडा मसाज 2 अंडों के पीले भाग में 2 चम्मच ऑलिव आयल मिलाकर पेस्ट बना लें। इसे 2 घंटे तक बालों में लगा रहने दें। फिर नहा लें। सीसम आयल 3-4 चम्मच सीसम आयल में एक चम्मच नींबू का रस मिला लें। इसे बालों और स्कैल्प पर लगाएं। इसके बाद गर्म तौलिये को सिर पर लपेटें और 5 मिनट तक भाप दें। इसके एक घंटे बाद नहा लें।
ENGLISH TRANSLATIONS	
Home Remedies For Dry Hair	Utility Desk: Many people experience dry hair problems in the winter. In such cases, hair appears dry and lifeless. However, we can address this issue through some home remedies. In this video, we will share a few such home remedies that, if used 3 to 4 times a week, can help reduce the problem of dry hair and make hair shiny. Olive Oil: Warm 3 to 4 tablespoons of olive oil. Massage it into your hair while it is lukewarm. Then wrap your hair with a slightly warm towel. After an hour, take a shower. Coconut Water: Mix 3 tablespoons of coconut water with 1 tablespoon of lemon juice. Apply it to your hair. Steam your hair for 5 minutes. Then, take a shower after an hour. Egg Massage: Make a paste by mixing 2 tablespoons of olive oil with the yolks of 2 eggs. Apply this paste to your hair and leave it on for 2 hours. Then, take a shower. Sesame Oil: Mix 3 to 4 tablespoons of sesame oil with 1 tablespoon of lemon juice. Apply it to your hair and scalp. Wrap your head with a warm towel and steam for 5 minutes. Take a shower after an hour.

Figure 29: An example of a query with its corresponding golden corpus from the CC News Retrieval Dataset

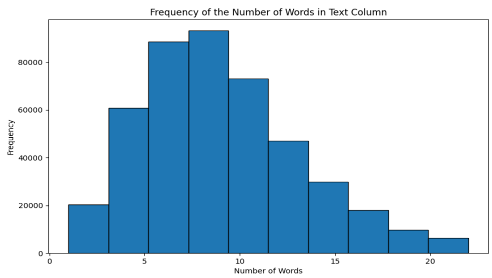


Figure 30: Distribution of the number of words in the queries of CC News Retrieval Dataset

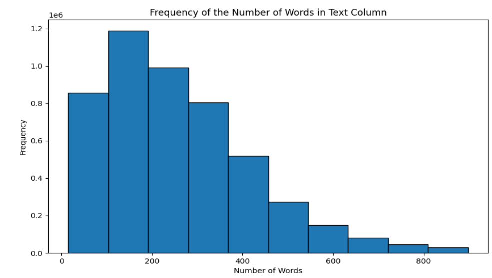


Figure 31: Distribution of Number of Words in corpus of CC News Retrieval Dataset

A.2.11 Sangraha-IR

1. **Task Definition:** Sangraha-IR introduces the task to retrieving relevant text documents given a question in Hindi. Subsection 3.2 talks in detail about the data creation process.
2. **Domain :** Miscellaneous

An example of a query with its corresponding golden corpus from the Sangraha-IR dataset has been provided in Figure 32

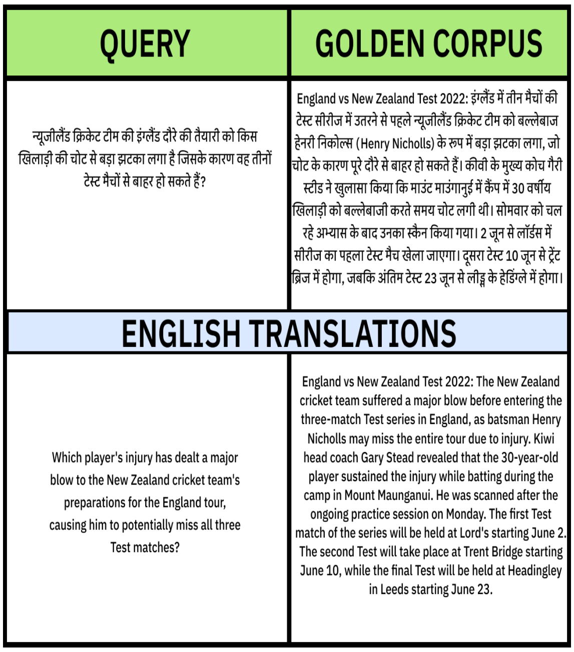


Figure 32: An example of a query with its corresponding golden corpus from the Sangraha-IR Dataset

Distribution of the number of words in the corpus and queries in the Sangraha-IR dataset has been shown in Figure 34 and Figure 33, respectively.

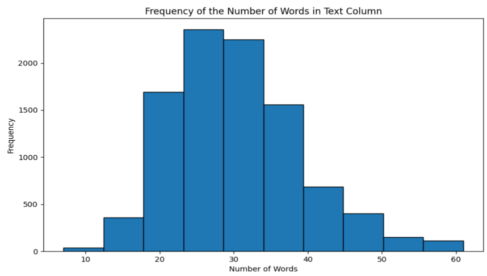


Figure 33: Distribution of the number of words in the queries of Sangraha-IR Dataset

A.2.12 MIRACL

1. **Task Definition:** MIRACL has been created from the Wikipedia dump of each language. Only plain text is considered, while images, etc., are omitted. The plain text is split up into multiple paragraphs, which act as the corpus. We consider the dev split of the Hindi subset of the original MIRACL in the Hindi-BEIR Benchmark.
2. **Domain :** Wikipedia

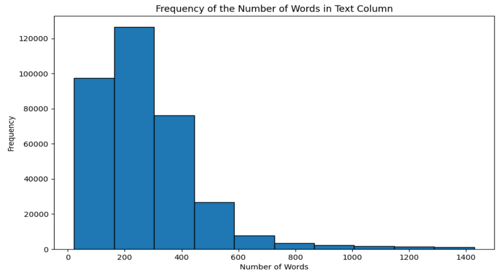


Figure 34: Distribution of Number of Words in corpus of Sangraha-IR Dataset

An example of query with its corresponding golden corpus from the MIRACL dataset has been provided in Figure 35

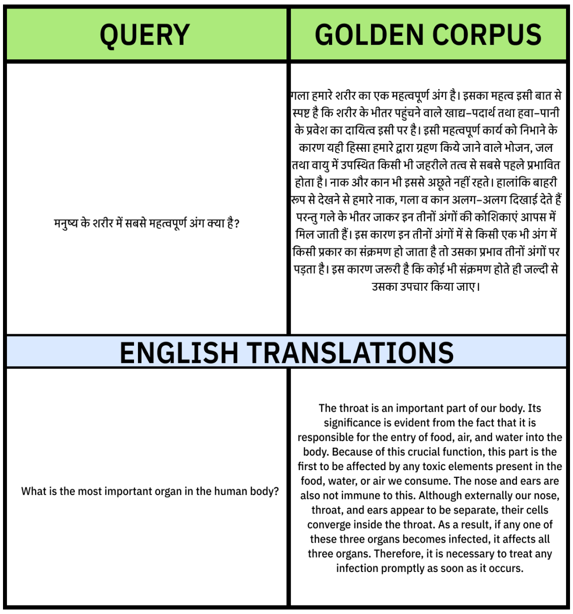


Figure 35: An example of a query with its corresponding golden corpus from the MIRACL Dataset

Distribution of the number of words in the corpus and queries in the MIRACL dataset has been shown in Figure 37 and Figure 36 respectively.

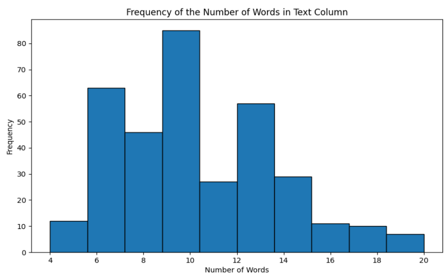


Figure 36: Distribution of the number of words in the queries of MIRACL Dataset

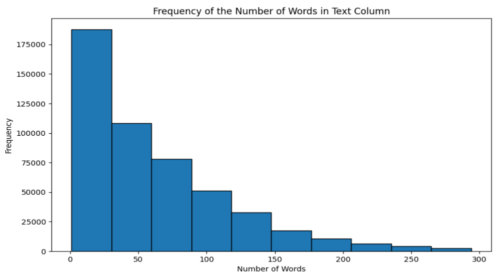


Figure 37: Distribution of Number of Words in the corpus of MIRACL Dataset

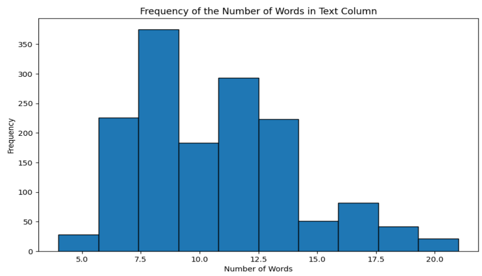


Figure 39: Distribution of the number of words in the queries of IndicQARetrieval Dataset

A.2.13 IndicQARetrieval

1. **Task Definition:** It is created by transforming the IndicQA dataset [12] to a retrieval dataset. The task is similar to NQ, where the model is expected to return the paragraph that contains the answer to the question when given a question.
2. **Domain :** Wikipedia

An example of query with its corresponding golden corpus from the IndicQARetrieval dataset has been provided in Figure 38

QUERY	GOLDEN CORPUS
'"नेपाल ने एक अच्छा दोस्त खो दिया है" यह बात किसने कही थी?'	"दलाई लामा ने अपनी संवेदना और प्रार्थना व्यक्त की और कलाम की मौत को "एक अपूरणीय क्षति" बुला, अपना दुख व्यक्त किया। उन्होंने यह भी कहा, "अनेक वर्षों में, मुझे कई अवसरों पर कलाम के साथ बातचीत करने का मौका मिला। वह एक महान वैज्ञानिक, शिक्षाविद और राजनेता ही नहीं, बल्कि वे एक वास्तविक सज्जन थे, और हमेशा मैंने उनकी सादगी और विनम्रता की प्रशंसा की है। मैंने सामान्य हितों के विषयों की एक विस्तृत श्रृंखला पर हमारी चर्चाओं का आनंद लिया, लेकिन विज्ञान, अध्यात्म और शिक्षा के साथ मुख्य रूप से हमारे बीच चिंतन किया जाता था। " दक्षिण एशियाई नेताओं ने अपनी संवेदना व्यक्त की और दिवंगत राजनेता की सराहना की। भूटान सरकार ने कलाम की मौत के शोक के लिए देश के झंडे को आधी ऊंचाई पर फहराने के लिए आदेश दिया, और श्रद्धांजलि में 1000 मक्खन के दीपक की भेंट किए।....SHORTENED DUE TO LENGTH
ENGLISH TRANSLATIONS	
'Who said "Nepal has lost a good friend"?	The Dalai Lama expressed his condolences and prayers, calling Kalam's death "an irreparable loss" and expressing his grief. He also said, "Over the years, I had several opportunities to interact with Kalam. He was not only a great scientist, educator, and statesman but also a truly noble person, and I always admired his simplicity and humility. I enjoyed our discussions on a wide range of common interests, primarily focused on science, spirituality, and education." South Asian leaders expressed their condolences and praised the late statesman. The Bhutan government ordered the national flag to be flown at half-mast in mourning for Kalam's death and offered 1000 butter lamps as a tribute....SHORTENED DUE TO LENGTH

Figure 38: An example of a query with its corresponding golden corpus from the IndicQARetrieval Dataset

Distribution of the number of words in the corpus and queries in the IndicQARetrieval dataset has been shown in Figure 40 and Figure 39 respectively.

[12] https://huggingface.co/datasets/ai4bharat/IndicQA

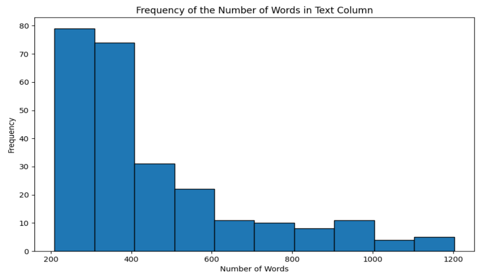


Figure 40: Distribution of Number of Words in the corpus of IndicQARetrieval Dataset

A.2.14 mMARCO

1. **Task Definition:** It is a multilingual version of the MSMARCO dataset. The dataset contains the translation of queries from Bing search logs with one text passage from various web sources annotated as relevant.
2. **Domain :** Miscellaneous

An example of a query with its corresponding golden corpus from the mMARCO dataset has been provided in Figure 41

QUERY	GOLDEN CORPUS
कार्ड भेजने की मोहर कितनी है	जून 2014 तक, यूनाइटेड स्टेट्स पोस्टल सर्विस का कहना है कि एक मानक आकार के पोस्टकार्ड को भेजने की लागत 34 सेंट है। इस दर पर, पोस्टकार्ड 6 इंच लंबा, 4 1/4 इंच ऊंचा और 0.16 इंच मोटा हो सकता है।
ENGLISH TRANSLATIONS	
How much is the stamp to send a card	As of June 2014, the United States Postal Service says the cost to send a standard-sized postcard is 34 cents. At this rate, the postcard can be 6 inches long, 4 1/4 inches high, and 0.16 inches thick.

Figure 41: An example of a query with its corresponding golden corpus from the mMARCO Dataset

Distribution of the number of words in the corpus and queries in the mMARCO dataset has been shown in Figure 43 and Figure 42 respectively.

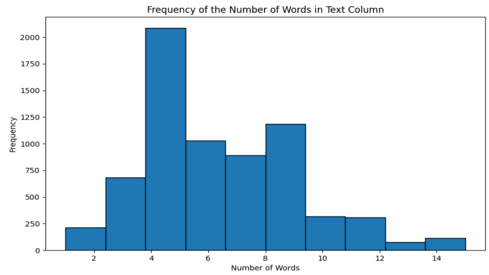


Figure 42: Distribution of the number of words in the queries of mMARCO Dataset

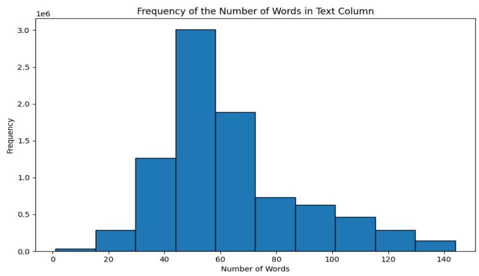


Figure 43: Distribution of Number of Words in the corpus of mMARCO Dataset

A.2.15 WikiPediaRetreival

1. **Task Definition:** It is similar to a question-answering task dataset like NQ, where given a query, the model is expected to retrieve a relevant article which answers the question. We have included the Hindi subset of WikiPediaRetrieval dataset [13]

2. **Domain :** Wikipedia

An example of a query with its corresponding golden corpus from the WikiPediaRetreival dataset has been provided in Figure 44

Distribution of the number of words in the corpus and queries in the WikiPediaRetreival dataset has been shown in Figure 46 and Figure 45 respectively.

A.3 Why choose IndicTrans2 over other available translation models?

Ans: Gala et al. (2023) clearly illustrate the superior performance of IndicTrans2 over other models

[13] https://huggingface.co/collections/ellamind/mmteb-6661723dc229e1da8e837cdf

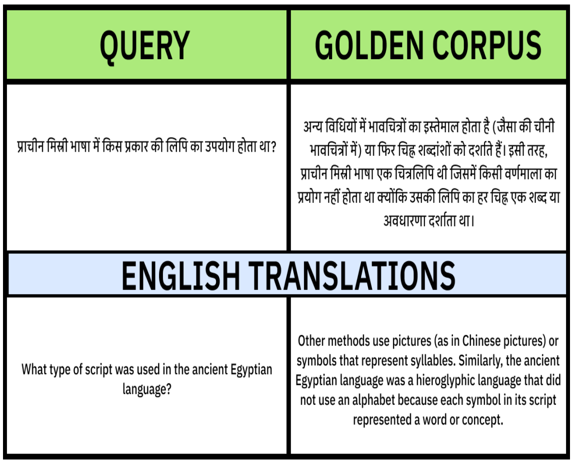


Figure 44: An example of a query with its corresponding golden corpus from the WikiPediaRetreival Dataset

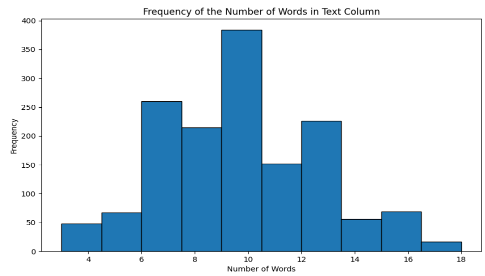


Figure 45: Distribution of the number of words in the queries of WikiPediaRetreival Dataset

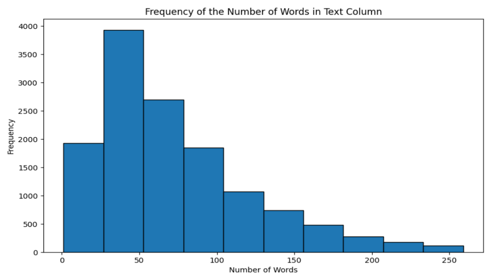


Figure 46: Distribution of Number of Words in the corpus of WikiPediaRetreival Dataset

and systems like NLLB and Google Translate for English to Hindi tasks. In our preliminary analysis on a subset of the BEIR datasets (Chrf scores show in Table 4), we also observed that IndicTrans2 outperformed alternative models, such as NLLB, in terms of translation quality.

A.4 Prompts used for Sangraha-IR Query Generation

The prompt that was used to obtain answers from the Gemini-Flash-1.5 Models is as follows

Given the above text, generate a question in

Dataset	IndicTrans2	NLLB
Arguana	56.30	37.12
NQ	76.13	60.92

Table 4: Chrf scores between the original English text and back translated english text by the respective model for 20,000 randomly selected samples from each dataset.

Hindi which can only be answered by the above passage. The question should be difficult with only semantic similarity and should not contain any lexical overlap; that is, do not directly use the exact phrases as shown in the above passage. You can use synonyms and make the question as complex as possible. Only return the relevant question and nothing else in the format given below : <QUESTION>: ...question in hindi.... < /QUESTION>

A.5 Ablation of LoRA parameters

Dataset Name	NLLB$_{1.3B(dist)}$-E5 (W/O LoRA, ST)	NLLB$_{1.3B(dist)}$-E5 (W LoRA, ST)
ArguAna	57.24	58.19
FiQA-2018	34.19	34.46
TREC-COVID	70.53	70.52
SCIDOCS	18.23	18.06
SciFact	64.36	65.23
Touché-2020	25.21	25.74
NQ	53.38	51.09
FEVER	71.04	72.71
Climate-FEVER	22.51	23.63
CC News Retrieval	31.91	29.47
MIRACL	52.96	49.76
IndicQARetrieval	62.10	61.81
mMARCO	34.03	31.78
WikiPediaRetrieval	86.20	84.36
Average	**48.84**	48.34

Table 5: NDCG@10 scores of NLLB-E5 model with and without LoRA adapters on E5 model on the Hindi-BEIR Benchmark.

Similar to Schmidt et al. (2024), we also experiment with the addition of LoRA adapters during training on top of the E5 model. The LoRA adapters were applied to the "query", "key", and "value" layers of the retrieval model head, with a rank of 32, an alpha parameter of 64, and a dropout rate of 0.05.

Table 5 presents a side-by-side comparison of the NLLB-E5 model with and without LoRA adapters applied to the E5 head. The results indicate a slight performance boost on certain datasets, particularly those derived from the original BEIR benchmark, which closely resembles the training data. However, it is important to note a significant drop in performance on non-BEIR datasets, such as CC-News Retrieval, MIRACL, IndicQA Retrieval, mMARCO, and Wikipedia Retrieval. We hypothesize that this drop could be due to the LoRA adapters causing the model to overfit on datasets in the style of Sentence Transformers, thereby reducing its generalization capabilities. The overall stronger performance of the model without LoRA adapters is further supported by its higher average NDCG@10 score, and thus, we decided to proceed with the "without LoRA" model setup.

A.6 Performance of NLLB Encoder

We trained the NLLB-Encoder using the same knowledge distillation approach as NLLB-E5 and evaluated its performance on subsets of both English and Hindi datasets from BEIR and Hindi-BEIR. The results, presented in Table 6, reveal a significant performance gap between NLLB-Encoder and both NLLB-E5 and mE5-Large in the retrieval task. These findings indicate that NLLB-Encoder alone is insufficient to act as a retrieval model and requires additional components, thereby motivating the development of NLLB-E5.

Model	Hindi ArguAna	Hindi FiQA-2018	IndicQARetrieval
NLLB(1.3B)-Encoder	44.45	17.54	42.98
NLLB-E5(1.3B)	**57.24**	**34.19**	61.94
mE5-Large	54.77	27.33	**67.11**

Table 6: NDCG@10 scores of NLLB-Encoder along with mE5-Large and NLLB-E5 on ArguAna and FiQA-2018 from Hindi-BEIR