

WiNCC: A Reproducible Pipeline for Constructing a Wikipedia-enriched Nepali Cross-lingual Corpus

Ahrnish P Dahal¹, Pushkar Baranwal², Amit Pundir³, Geetika Jain Saxena³,
Sanjeev Singh²

Cluster Innovation Centre, University of Delhi¹
Institute of Informatics and Communication, University of Delhi²
Maharaja Agrasen College, University of Delhi³

Abstract

WiNCC is a reproducible and automatic construction pipeline for building English–Nepali cross-lingual retrieval corpus that is difficulty-aware, relying on public Wikipedia and Wikidata.

Part 1 reproduces a CLIRMatrix-style baseline. Nepali Wikipedia was crawled, articles resolved to a Wikidata QID then intersected with corresponding Wikidata sitelinks to obtain Nepali–English query pairs and related English document corpus. These were sampled to gold pairs and padded with random English distractors into a document baseline corpus.

Part 2 adds a data-enrichment stage to Part 1 with multi-query, class-matched hard negatives. We compare the produced corpus against existing cross-lingual retrieval resources, run cross-lingual retrieval baselines with three distinct multilingual embedding models (BGE-M3, multilingual-E5-large, and Qwen3-Embedding-4B), analyze slice-wise quality-metadata and query-type-wise field metrics supporting the corpus’s use as a difficulty-aware benchmark.

Through retrieval tasks on our corpus, we also find (1) a 27.8–41.5 point NDCG@10 drop from English to Nepali, (2) descriptions are 5–6x harder than titles, and (3) hard negatives are 2x stickier than random distractors.

1 Introduction

Building evaluation resources for low-resource, cross-lingual retrieval by hand requires bilingual annotators or expensive machine-translation and labelling step. For this work, a retrieval corpus consisting of English Wikipedia documents was crawled; the query language varies (obtained from title queries for English and Nepali). We follow the automatic language-linking methodology of CLIRMatrix [Sun and Duh, 2020] and its sentence-centric predecessor WikiCLIR [Sasaki et al., 2018] to construct an English–Nepali query bucket directly from Wikipedia’s per-language editions and

Wikidata’s cross-lingual entity graph, with no need for translation or human relevance judgment. The corpus construction procedure is fully automatic from public APIs, and reported stage-by-stage, with an aim to make it directly reproducible. Part 2 (also referred to as v2) adds a data-enrichment and quality-annotation layer (§4) to the Part 1 output, aligning with Web-as-Corpus call for work on data enrichment and quality.

Concretely, we contribute:

- (1) Reproducible corpus creation pipeline (§3);
- (2) Enrichment and hard-negative mining (§4);
- (3) Cross-lingual retrieval baselines (§5);
- (4) Metadata-sliced evaluation (§6).

Extended per-slice tables and all supporting visualizations are provided in Appendix A.

2 Related Work

This section discusses the relevant literature and previous works referenced and also compares them against our work.

Web corpus construction and filtering. OSCAR [Suárez et al., 2019] worked on a multi-threaded pipeline to clean and classify Common Crawl by language built especially for medium-resource infrastructures; CCNet [Wenzek et al., 2020] extracted massive high-quality monolingual datasets from Common Crawl for a variety of languages; and ParaCrawl [Bañón et al., 2020] mines web-scale parallel corpora by crawling the web.

Automatic cross-lingual dataset construction. CLIRMatrix [Sun and Duh, 2020] used Wikipedia’s inter-language structure, via Wikidata [Vrandečić and Krötzsch, 2014] to construct a collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval. WikiCLIR [Sasaki et al., 2018] is the precedent that automatically extracted German-English retrieval dataset for Cross-Language Information Retrieval. WikiMatrix [Schwenk et al., 2021] mines Wikipedia for

parallel sentences rather than paired queries and documents.

Multilingual retrieval benchmarks. Widely used multilingual retrieval benchmarks MIRACL [Zhang et al., 2023] and Mr. TyDi [Zhang et al., 2021] rely on human relevance judgments and their language set does not include Nepali. Cross-lingual QA datasets such as XOR-QA [Asai et al., 2021], TyDi QA [Clark et al., 2020], and MKQA [Longpre et al., 2021] target human-authored-QA problem. Table 1 compares WiNCC against these existing resources.

Multilingual representation and dense retrieval. XLM-R [Conneau et al., 2020] established pretraining multilingual language models at scale improves performance for a wide range of cross-lingual transfer tasks; SentenceBERT [Reimers and Gurevych, 2019] showed that Siamese fine-tuning turns such encoders into fast, comparable sentence embeddings, and LaBSE [Feng et al., 2022] extended this to a language-agnostic objective. For retrieval, Dense Passage Retrieval [Karpukhin et al., 2020] proved that dual-encoder dense retrieval can outperform sparse lexical methods for open-domain QA, later extended multilingually. BGE-M3 [Chen et al., 2024], multilingual-E5-large [Wang et al., 2024], and Qwen3-Embedding [Zhang et al., 2025] represent the current generation of broad multilingual embedding models built on this lineage; we use all three only as external instruments for retrieval baselines.

Dense retrieval and hard-negative mining. A separate line of work improves dense retrievers: ANCE [Xiong et al., 2021] applies nearest neighbour negative contrastive learning for dense retrieval and RocketQA [Qu et al., 2021] works on cross-batch negatives; ColBERT [Khattab and Zaharia, 2020] implements late interaction in the scoring stage to make finer distinctions between close candidates; and Contriever [Izacard et al., 2022] mines hard negatives entirely unsupervised via contrastive pretraining.

3 Part 1: Baseline Corpus Construction

This section discusses the construction process and stage-wise counts of the various components in our baseline dataset.

Crawl and Wikidata linking. The pipeline crawls Nepali Wikipedia via MediaWiki endpoints. A full crawl checked **42,434** Nepali articles and resolved **30,689** to a Wikidata QID. Each QID was

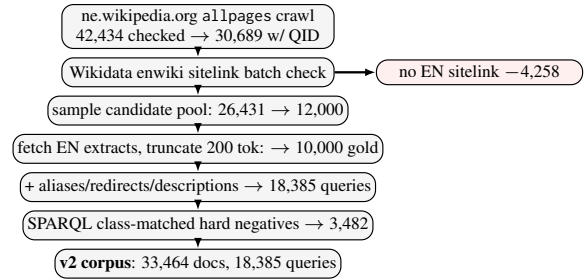


Figure 1: End-to-end pipeline (Part 1 + Part 2), with per-stage counts.

batch-queried against Wikidata (in batches of 50), yielding **26,431** Nepali–English pairs (~86% of QID-resolved articles have an English counterpart).

Sampling and gold document collection. From the 26,431 pairs we sampled 12,000 pairs (random_state=42) as a safe pool (in case of rate limit failures during extraction) against a set target of 10,000 gold pairs. English extracts were fetched in batches of 20, truncated to 200 whitespace tokens, following CLIRMatrix [Sun and Duh, 2020], till 10,000 gold documents were collected. (Min/Mean/Max = 4 / 114.9 / 200 tokens).

Baseline distractor construction. Distractors were drawn via list=random, deduplicated against gold titles and required to have ≥ 20 tokens, truncated identically, until exactly 20,000 were collected (Min/Mean/Max = 20 / 77.0 / 200 tokens). The baseline corpus contains **30,000** documents (10,000 gold + 20,000 distractors).

4 Part 2: Multi-Query, Hard Negatives, and Metadata

This section discusses the procedure and numbers associated with the Part 2 dataset. Part 2 expands the dataset across three key dimensions:

Expanded Multi-Query Retrieval. To reflect realistic search behaviours, three additional Nepali queries were mapped to gold entities via wbgetentities: Wikidata aliases (1,035), Wikipedia redirects (4,277), and Wikidata descriptions (3,073). Combined with 10,000 original titles, this yields **18,385 query formulations** across 10,000 gold documents.

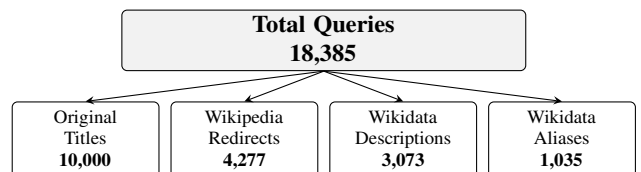


Figure 2: Distribution of the 18,385 total queries across their source categories.

Class-Based Hard Negatives. Topically close

Corpus	#Langs	Nepali	Construction	Relevance	Hard neg.	Multi-query	Metadata
WikiCLIR [Sasaki et al., 2018]	1 pair	no	Wikipedia-derived	binary/graded, auto	no	no (first-sentence query)	no
CLIRMatrix [Sun and Duh, 2020]	~139	incl., unaudited	Sitelinks	binary, auto	no	no (title only)	no
MIRACL [Zhang et al., 2023]	18	no	Human judgments	graded	implicit (pooled)	no	no
Mr. TyDi [Zhang et al., 2021]	11	no	Human judgments	binary	implicit (pooled)	no	no
WikiMatrix [Schwenk et al., 2021]	1620 pairs	sentence-level	Margin mining	n/a (bitext)	n/a	n/a	no
XOR-QA [Asai et al., 2021]	7	no	Human-authored	binary	no	no	no
WiNCC (ours)	1 pair	dedicated	Wikidata/WP	binary, auto	SPARQL class-matched	1.84 qry/entity	5 fields

Table 1: WiNCC vs. existing cross-lingual retrieval and bitext resources. “Multi-query” = more than one query form per gold document; “Metadata” = attached per-instance difficulty/quality fields.

distractor documents were mined using SPARQL queries against primary Wikidata classes. Filtering for documents with ≥ 20 tokens yielded 3,482 **hard negatives** (38.5% yield) from 9,072 initial candidate titles.

Per-Instance Metadata. Each instance is enriched with four metadata fields: `alias_count`, `ambiguity_level`¹, `entity_category`, and `lexical_overlap`.

Part 2 corpus assembly. The final v2 corpus merges gold documents, hard negatives, and baseline random distractors: 10,000 gold, 3,482 hard negatives, 19,982 random distractors (18 documents fell through during rate limiting), for a total of **51,849 entities : 33,464 documents** against **18,385 query formulations** (Table 2).

Component	Part 1	Part 2 (v2)
Gold documents	10,000	10,000
Query formulations	10,000	18,385
Hard negatives	–	3,482
Random distractors	20,000	19,982
Total corpus size	30,000	51,849

Table 2: Component counts, baseline vs. v2 corpora.

Dataset Access and Structure

The complete WiNCC corpus, including both baseline and enriched variants along with pipeline scripts are available at [WiNCC](#).

The data release is organized into two directories, corresponding to the construction stages described in §3 and §4

- Part 1/: Contains the baseline 30,000-document corpus.
 - `query_pairs_final.csv`: The 10,000 Nepali–English title query pairs.
 - `gold_docs_truncated.json`: The 10,000 gold English Wikipedia extracts, truncated to 200 whitespace tokens.
 - `distractor_docs.json`: The 20,000 random English distractor documents.

¹Ambiguity is overwhelmingly low: 18,336/18,385 query rows (99.7 %) fall in the low bucket, with only 43 medium and 6 high rows.

- `corpus_full.csv`: The complete baseline corpus, merging gold and distractor documents with their binary relevance labels.
- Part 2/: Contains the enriched v2 corpus with multi-query formulations and hard negatives.
 - `multi_query_benchmark_enriched.csv`: The full 18,385 query formulations (titles, redirects, aliases, and descriptions) with attached per-instance metadata (ambiguity, category, lexical overlap, etc.).
 - `hard_negative_titles_by_gold.json`: Mapping of the 3,482 class-matched hard-negative document titles to their corresponding gold entities.
 - `corpus_full_v2.csv`: The final 33,464-document corpus, combining gold documents, hard negatives, and random distractors with their respective `doc_type` labels.
 - `wikidata_entity_details.json`: Cached Wikidata entity metadata (aliases, descriptions, and QID properties) used during the enrichment stage.

This results in a total of 51,849 entities. (inclusive of query formulations). The repository also includes `wincc_creation.ipynb`, a notebook for the complete pipeline of corpus creation. All files are provided in plain CSV and JSON formats for ease of use with standard retrieval toolkits.

5 Cross-Lingual Retrieval Baselines

All retrieval experiments use the same English document corpus constructed in §3 and §4. We vary only the query language, comparing retrieval task with English query formulations against Nepali query formulations. We report retrieval baselines from three architecturally distinct multilingual embedding models – BGE-M3 [Chen et al., 2024], multilingual-E5-large [Wang et al., 2024], and Qwen3-Embedding-4B [Zhang et al., 2025] – encoding all v2-corpus documents and both English and Nepali queries, retrieving (mono- and cross-lingual) with FAISS index [Johnson et al., 2019].

mE5 uses its "query: "/"passage:" prefix convention; Qwen3-Embedding uses its query-instruction prompt; BGE-M3 uses symmetric encoding.

Model	Query	R@1	R@10	R@100	NDCG@10
BGE-M3	EN	0.945	0.990	0.995	0.971
	NE	0.502	0.706	0.831	0.603
mE5-large	EN	0.950	0.991	0.995	0.973
	NE	0.449	0.671	0.820	0.558
Qwen3-Emb-4B	EN	0.971	0.993	0.996	0.984
	NE	0.627	0.783	0.878	0.706

Table 3: Overall retrieval metrics by query language, three models. All documents in English. EN = English query, NE = Nepali query. Full per-slice breakdowns are in Appendix A.

All three models are near-ceiling on English queries and show a consistent drop on Nepali queries. Qwen3-Embedding-4B is consistently strongest for English and Nepali queries, mE5-large consistently weakest for Nepali queries. This convergence across three independent models confirms both that gold labels are correct and recoverable, and that the difficulty effect is a property of the corpus’s query-language design.

6 Metrics Sliced by Query Metadata

Query-metadata and their associated metrics were a revealing study. Queries were divided in the following ways.

By query type. Titles are consistently easiest (NDCG@10 0.66–0.81), redirects and aliases close behind, and descriptions by far hardest (0.16–0.26): as a paraphrase requires semantic search tasks beyond entity lookup, so retrieval degrades. Identical results ordering for all three models highlights a corpus property, not a single model’s weakness.

By ambiguity. (Fig. 3c.) Higher-ambiguity queries score better (0.60–0.71 low vs. 0.74–0.89 high). This is not conclusive as the ambiguity bucket was very tiny (n=6) dominated by short, distinct proper names with easy string matches.

By entity category. Taxon entities collapse to NDCG@10 0.08–0.12 for every model, roughly one-sixth the average, suspected due to asymmetric pre-training of models on scientific nomenclature. Administrative and human categories sit close to the pooled mean.

By lexical overlap. Devanagari queries vs. Latin titles show uniformly low character overlap (0.04–0.08), yet the some-overlap slice (≥ 0.05 , n=10,830, mostly transliterated proper nouns) still beats the zero-overlap slice (n=7,555) by 6–13 Recall@1 points for every model.

Model	Form	R@1	R@10	R@100	NDCG@10
BGE-M3	Title	0.600	0.796	0.879	0.699
	Redirect	0.549	0.772	0.867	0.660
	Alias	0.503	0.749	0.861	0.626
	Desc.	0.120	0.306	0.612	0.203
mE5-large	Title	0.543	0.766	0.870	0.655
	Redirect	0.488	0.732	0.854	0.608
	Alias	0.454	0.712	0.840	0.579
	Desc.	0.085	0.262	0.602	0.162
Qwen3-Emb-4B	Title	0.745	0.868	0.909	0.810
	Redirect	0.686	0.841	0.894	0.769
	Alias	0.648	0.846	0.911	0.754
	Desc.	0.152	0.403	0.745	0.265

Table 4: Retrieval metrics by query formulation type. Query counts: Title=10,000, Redirect=4,277, Alias=1,035, Desc.=3,073.

Further breakdowns in Table 5, heatmaps in Appendix A, Figs. 3–4).

7 Do Hard Negatives Actually Confuse the Model More than Random Distractors?

This section highlights the methods and subsequent findings related to hard-negatives introduced in §4.

Method. For each of the 10,000 gold entities, we tag each top-10 retrieved document by doc_type and compare its observed share of top-10 slots against its corpus base share (gold 29.9%, hard negative 10.4%, random distractor 59.7%). We also check, per gold entity, whether mined hard negative outranks gold. We validate difficulty via model behavior rather than human judgment, consistent with prior model-agnostic hard-negative work.

Result. Random distractors appear in the top-10 results far less often than their share of the corpus (only $0.31\text{--}0.58\times$ their base rate) for every model. Hard negatives are also under-represented, but much less severely ($0.69\text{--}0.87\times$ their base rate). Relative to corpus frequency, a hard negative is therefore roughly twice as likely to reach the top-10 as a random distractor. Gold documents are over-represented, as expected (lift $1.88\text{--}2.50\times$; full numbers in Table 6, Appendix A, Fig. 4c–d).

Despite this increased difficulty, hard negatives almost never outrank the correct gold document at rank 1 (only 0.14–0.41% of cases; Appendix Fig. 5a). The same $\sim 2\times$ relative-lift pattern holds across all three models, which use different pre-training data, objectives, and prompting conventions (§5). This consistency indicates that the hard negatives bring a genuine topical-confusability signal, in the same spirit as model-agnostic hard-negative validation in prior work [Xiong et al., 2021, Qu et al., 2021].

8 Conclusion

We presented a reproducible, fully-automatic pipeline for constructing English–Nepali cross-lingual retrieval corpora from public Wikipedia and Wikidata APIs, instantiated as WiNCC: 33,464 documents, 18,385 query formulations across 10,000 gold entities, and 3,482 class-matched hard negatives. Beyond construction, we introduced a metadata-sliced evaluation methodology that pinpoints exactly where cross-lingual difficulty concentrates (description-type queries, taxon-category entities), using three independently-trained multilingual embedding models to verify this. Results also show roughly double the retrieval lift of random distractors without ever outranking gold documents. Together these numbers support the corpus’s use as a genuine difficulty-aware cross-lingual benchmark and the pipeline’s use as a reusable template for other low-resource language pairs.

9 Limitations and Discussions

- **Structure-based labeling.** Language labels obtained faithfully from Wiki language site partitioning, and relevance is simplified to binary, not graded.
- **Rate-limiting affected hard-negative coverage.** Rate-limiting during Part 2 curation affected the extraction of queries and dropped the attainment to 19,982 distractors (§4 and Table 2) than the aimed 20,000 (as is in Part 1).
- **Single language pair, non-comparative baselines.** Only English–Nepali language pair studied; §5 numbers are reference baselines only.
- **Comparison table.** Table 1 includes retrieval, bitext as well as QA onto shared axes for orientation only, not like-for-like equivalence comparison.

The class-matched hard negatives, proved in §7 to carry a genuine, and consistent confusability signal help in difficulty-controlled evaluation. The metadata-sliced methodology of §6 supports slice-based error analysis. Hard-negative pools can serve as contrastive training pairs [Karpukhin et al., 2020, Izacard et al., 2022], mined independently of any model’s performance, unlike iterative model-dependent mining approaches [Xiong et al., 2021]. As Table 1 shows, WiNCC is, to our knowledge, the only automatic CLIR resource combining a dedicated Nepali subset, class-matched hard

negatives, multiple query formulations per entity, as well as per-instance quality metadata simultaneously; the pipeline itself is reusable for other Wikipedia/Wikidata language pairs facing the same low-resource evaluation gap. The baselines and metrics are reference numbers attached to this release, not the paper’s main claim, and are a cheaper but weaker form of evidence, complementary to annotation by humans supporting MIRACL and Mr. TyDi.

References

- Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. Xor qa: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 547–564.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarriás, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4555–4567.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research (TMLR)*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.
- Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mlqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 5835–5847.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. 2018. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1351–1361.
- Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the 7th Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, 2019, Cardiff, Wales, UK, pages 9–16.
- Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4003–4012.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*.
- 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 (MRL)*, pages 127–137.
- 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.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Extended Results and Visualizations

This appendix gives the full metadata-sliced breakdowns and supporting visualizations summarized in §6–§7, for all three baseline models, enlarged here for readability.

A.1 Query-Formulation-Type and Ambiguity Slices

Corresponds to the query-type and ambiguity discussion in §6. Figure 2(b)-(c) covers this.

A.2 Entity-Category and Lexical-Overlap Slices

Corresponds to the entity-category and lexical-overlap discussion in §6. Panels (a)–(b) of Figure 4 below cover this subsection; panels (c)–(d) are referenced again in §A.3.

A.3 Hard-Negative Validation

Corresponds to §7. Panels (c)–(d) of Figure 4 show the same top-10 composition and lift results tabulated exactly in Table 6 below.

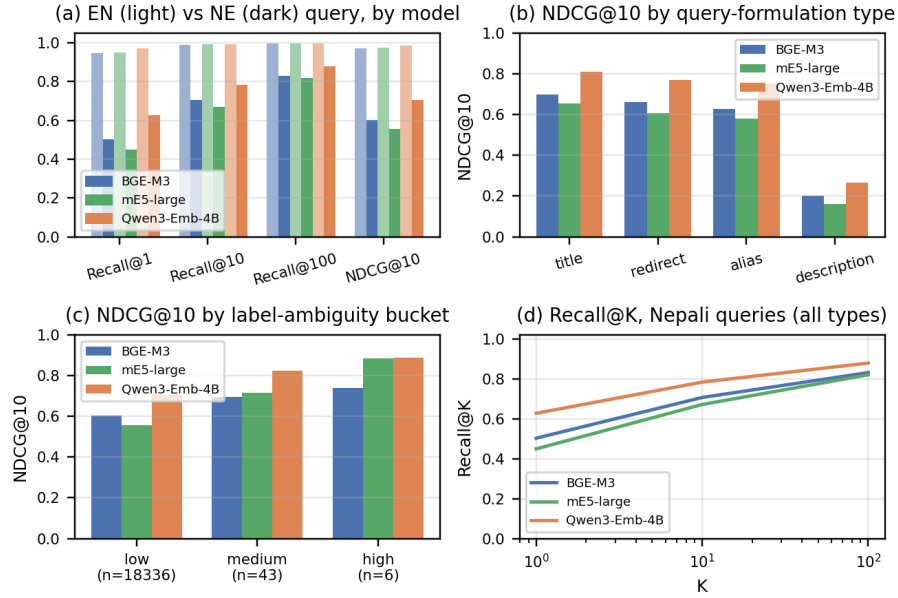


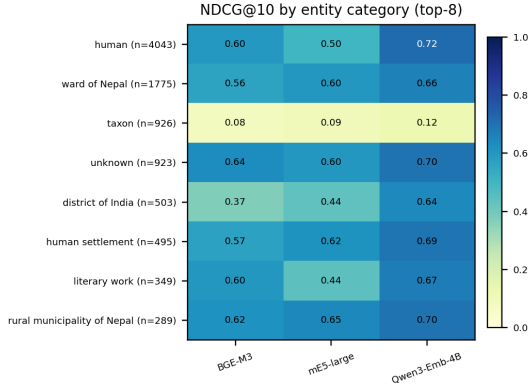
Figure 3: (a) EN (light) vs. NE (dark) pooled metrics by model; (b) NDCG@10 by query-formulation type; (c) NDCG@10 by ambiguity bucket; (d) Recall@K curve, Nepali queries (anchored at exact R@1/R@10/R@100).

Model	Metric	human	ward of NP	taxon	unknown	dist. of India	human settl.	lit. work
BGE-M3	R@1	0.502	0.467	0.054	0.536	0.239	0.465	0.513
	NDCG@10	0.599	0.564	0.081	0.635	0.371	0.568	0.604
mE5-large	R@1	0.388	0.518	0.064	0.504	0.322	0.511	0.335
	NDCG@10	0.498	0.603	0.094	0.604	0.435	0.615	0.443
Qwen3-Emb-4B	R@1	0.644	0.579	0.093	0.626	0.517	0.604	0.602
	NDCG@10	0.715	0.663	0.123	0.699	0.644	0.694	0.672

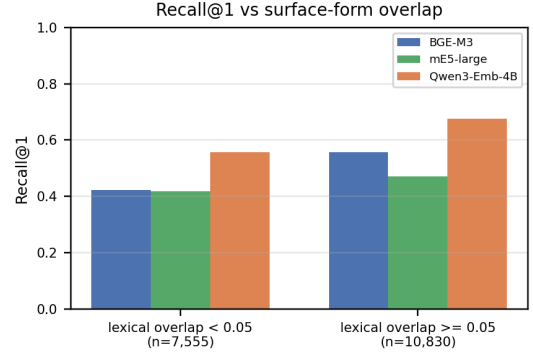
Table 5: Retrieval metrics sliced by entity category (top-7 of top-8 shown; rural municipality of Nepal, n=289, omitted for space – see Figure 4a for the full 8-category heatmap including it).

Model	Type	Obs. share	Base rate	Lift
BGE-M3	gold	0.563	0.299	1.88×
	hard neg.	0.091	0.104	0.87×
	random	0.346	0.597	0.58×
mE5-large	gold	0.746	0.299	2.50×
	hard neg.	0.072	0.104	0.69×
	random	0.182	0.597	0.31×
Qwen3-4B	gold	0.674	0.299	2.26×
	hard neg.	0.088	0.104	0.84×
	random	0.238	0.597	0.40×

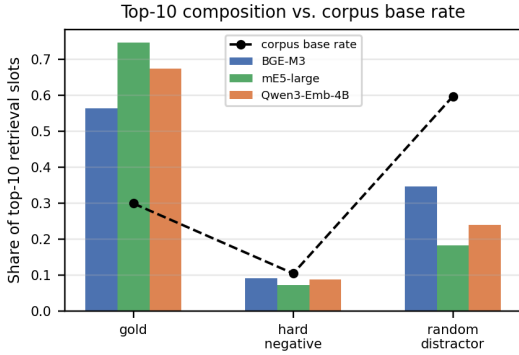
Table 6: Top-10 retrieval-slot composition vs. corpus base rate, by document type (title-query rows, n=10,000 per model). Lift = observed share / base rate.



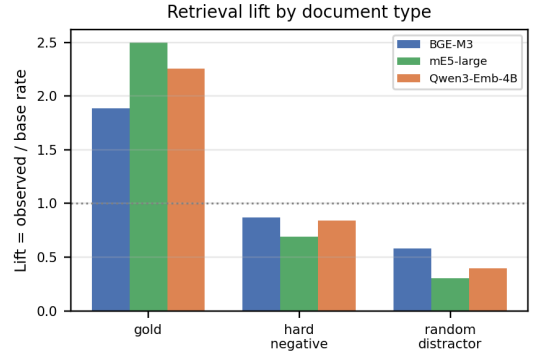
(a) NDCG@10 heatmap, top-8 entity categories $\times$ 3 models.



(b) Recall@1, zero- vs. some-lexical-overlap slices.

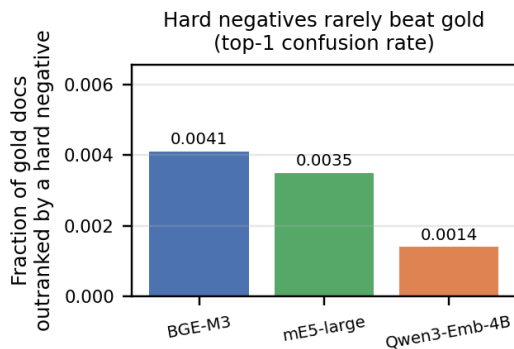


(c) Top-10 slot composition vs. corpus base rate.

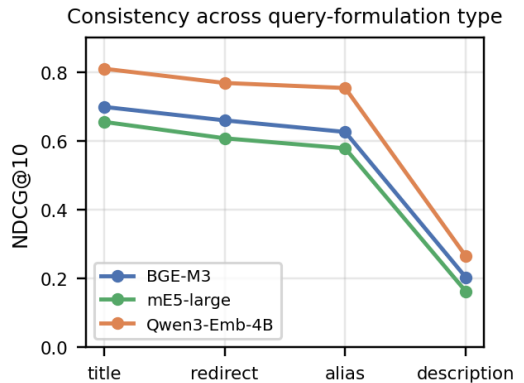


(d) Lift ratio (observed/base) by document type.

Figure 4: (a) NDCG@10 heatmap, top-8 entity categories $\times$ 3 models; (b) Recall@1, zero- vs. some-lexical-overlap slices; (c) top-10 slot composition vs. corpus base rate; (d) lift ratio (observed/base) by document type.



(a) Fraction of gold documents where a mined hard negative outranks gold at rank 1.



(b) NDCG@10 by query type, three models overlaid: an identical difficulty ordering (title > redirect > alias $\gg$ description) across architectures.

Figure 5: Hard-negative outranking rate (left) and cross-model query-type consistency (right).