Hybrid Sparse-Neural Fusion for Passage Retrieval

Georgios Arampatzis*
Avi Arampatzis

Democritus University of Thrace, Dept. of Electrical and Computer Engr., Xanthi, Greece

{geoaramp,avi}@ee.duth.gr

Abstract

This paper studies multilingual information retrieval (MLIR) and report generation under retrieval-augmented evaluation settings, with an emphasis on robustness, reproducibility, and interpretability. We focus on efficient and lightweight transformer-based cross-encoder architectures for passage re-ranking in a multilingual retrieval scenario.

Our approach follows a two-stage retrieval framework. In the first stage, BM25 is used for initial candidate selection, while in the second stage lightweight transformer-based cross-encoders (MiniLM, ELECTRA, and TinyBERT) are applied for passage re-ranking. To integrate multiple model predictions, we employ Reciprocal Rank Fusion (RRF), enabling robust aggregation across heterogeneous ranking signals.

Unlike multilingual fine-tuning or agent-based approaches, our system relies exclusively on English-only cross-encoders applied in a zero-shot setting. This design allows us to assess the cross-lingual generalization capacity of compact re-ranking models under strict efficiency and reproducibility constraints.

Experimental results on the validation phase show that, while our runs do not match the absolute effectiveness of large-scale multi-agent or generative systems, they achieve stable and interpretable performance given their lightweight architecture. Overall, the findings highlight the trade-off between retrieval effectiveness and computational efficiency, and demonstrate that compact re-ranking architectures combined with simple fusion strategies remain viable baselines for multilingual retrieval in low-resource and efficiency-constrained settings.

Keywords: Multilingual Information Retrieval, Hybrid Retrieval, BM25, Neural Re-ranking, Cross-Encoders, Reciprocal Rank Fusion

1 Introduction

The 2025 RAGTIME Track introduced new benchmarks for multilingual information retrieval (MLIR) and report generation, aiming to promote robustness, reproducibility, and interpretability in retrieval-oriented evaluation settings. In this work, the DUTH team (Democritus University of Thrace) presents its participation in the MLIR subtask, focusing on effective and lightweight transformer-based cross-encoder architectures for re-ranking candidate passages.

Multilingual information retrieval remains a challenging problem due to linguistic diversity, data scarcity, and domain-specific variation across languages. While dense retrieval models have demonstrated strong cross-lingual transfer capabilities [1, 2], their computational cost often limits their applicability in large-scale evaluation settings. As a result, hybrid pipelines that combine sparse retrievers such as BM25 [3] with neural cross-encoders [4] have emerged as practical and effective solutions, balancing efficiency and retrieval effectiveness.

Our approach follows a two-stage retrieval framework. In the first stage, we apply BM25, implemented via the Whoosh indexer, to retrieve an initial set of 1000 candidate passages per topic. In the second stage, multiple transformer-based cross-encoders (MiniLM-L6, MiniLM-L12, ELECTRA, and TinyBERT) are employed exclusively for passage re-ranking. To integrate predictions from individual models, we adopt Reciprocal Rank Fusion (RRF) [5], enabling robust aggregation across model variations.

Unlike fully multilingual systems trained on parallel corpora, our pipeline relies on English-only cross-encoders applied in a zero-shot setting, allowing us to assess their cross-lingual generalization capacity within the MLIR framework. Despite this simplicity, preliminary results on the validation phase indicate consistent retrieval effectiveness across English and translated document collections.

This paper describes the retrieval pipeline, model configurations, and system runs submitted to the TREC RAG-

*Corresponding author.

TIME 2025 MLIR track. Our findings highlight the trade-off between model efficiency and cross-lingual robustness in low-resource evaluation contexts.

Recent advances in lightweight transformer architectures and hybrid retrieval pipelines motivate a closer examination of their effectiveness in multilingual passage retrieval. Building on this perspective, the present work positions our system within the broader RAGTIME 2025 evaluation framework and outlines its methodological contributions. Section 2 reviews related work on sparse, dense, and hybrid neural retrieval approaches, as well as multilingual transfer techniques. Section 3 details the datasets, two-stage retrieval pipeline, model configurations, and fusion mechanisms used across our submissions. Section 4 reports the official NIST evaluation results and provides an analysis of performance trends across BM25, cross-encoder re-ranking, and hybrid RRF-based fusion. Finally, Section 5 summarizes the main findings and discusses future research directions, including multilingual fine-tuning, adaptive fusion strategies, and integration with retrieval-augmented generation pipelines.

2 Related Work

Multilingual information retrieval and cross-lingual transfer have been extensively studied in recent years, motivated by the increasing availability of multilingual corpora and the demand for retrieval systems that generalize across languages. Early work demonstrated that multilingual sentence and document embedding models can capture cross-lingual semantic similarity with limited supervision, enabling zero-shot or weakly supervised retrieval across languages [1, 2]. Despite their effectiveness, fully dense retrieval approaches often require substantial computational resources for indexing and inference, which can limit their applicability in large-scale or efficiency-constrained evaluation settings.

To address these limitations, hybrid retrieval architectures that combine sparse lexical retrieval with neural re-ranking have become increasingly popular. Sparse models such as BM25 [3] remain strong and widely used baselines due to their efficiency, interpretability, and robustness across domains. Neural cross-encoders [4], when applied as a second-stage re-ranking component, have been shown to substantially improve early ranking quality by modeling fine-grained query–document interactions. This two-stage paradigm allows systems to balance effectiveness and efficiency while retaining transparency in the retrieval process.

Beyond individual retrieval models, rank fusion methods have been explored as a means of integrating complementary retrieval signals. Late-fusion techniques such as Reciprocal Rank Fusion (RRF) [5] have demonstrated strong robustness across heterogeneous retrieval pipelines, particularly in TREC-style evaluation settings. By aggregating ranked lists rather than raw scores, RRF reduces sensitivity to score normalization and model-specific biases, making it well suited for combining sparse and neural re-ranking outputs.

Recent shared-task evaluations, including TREC benchmarks, have increasingly emphasized reproducibility, efficiency, and interpretability alongside raw effectiveness. Within this context, lightweight transformer architectures and simple fusion strategies provide a practical compromise between retrieval quality and computational cost, especially for multilingual and low-resource scenarios where extensive fine-tuning or large-scale dense indexing may not be feasible.

Beyond prior work in multilingual retrieval, the present study builds upon the DUTH team's previous research on lightweight transformer-based systems and hybrid evaluation frameworks. In earlier shared-task participations such as SemEval–2023 [6], EXIST–2025 [7], JOKER–2025 [8], and SimpleText–2025 [9], we developed modular and reproducible pipelines for tasks including affective analysis, bias detection, text simplification, and hallucination evaluation. These efforts emphasized interpretability, efficiency, and open-weight experimentation, which directly informed the retrieval design choices and fusion strategies adopted in the present work within the TREC RAGTIME 2025 MLIR framework.

3 System Description

Describe the datasets, indexes, topics/queries, and any external resources (including LLMs, pretraining corpora). Mention license and availability where applicable.

3.1 Dataset

The experiments conducted for this work were based on the RAGTIME Track dataset introduced in TREC 2025 [10]. The task is framed as a passage retrieval problem, where systems are required to rank a large collection of passages given a set of user queries. While the track includes multilingual data, our submission focused on the English subset for the MLIR (Multilingual Information Retrieval) task.

The English corpus is derived from the publicly available multilingual collection built upon the MS MARCO passage dataset [11] and extended to multiple languages using machine translation and alignment techniques, as described in [12]. Each passage is associated with a unique identifier and normalized metadata to facilitate consistent evaluation across languages and retrieval models.

For training and evaluation, the official topic sets provided by the RAGTIME organizers were used. The training split includes a mixture of queries sampled from previous TREC tracks and additional synthetic queries generated via large language models. The evaluation queries (topics 1001–1055) were used to generate the official run files submitted to TREC. Each query is associated with approximately 1,000 ranked passages per system run, following the standard TREC output format [13].

To ensure comparability across approaches, all text was tokenized and normalized using the same preprocessing pipeline adopted by the Hugging Face Transformers framework [14]. The dataset was stored locally in JSONL format, while the final submissions were produced in TREC's `.txt` run file format, consisting of lines of the form:

```
<topic-id> Q0 <doc-id>
<rank> <score> <run-tag>
```

where `<run-tag>` corresponds to the system identifier (e.g., `duth-mlir-rrf`).

Overall, the dataset provided a robust foundation for evaluating sparse and neural retrieval models under controlled, multilingual conditions, enabling fair benchmarking and fusion analysis of different cross-encoder architectures.

3.2 Models

The retrieval systems developed for the TREC 2025 RAGTIME Multilingual Information Retrieval (MLIR) task employed a combination of sparse lexical retrieval and neural cross-encoder models for second-stage re-ranking. All models were implemented using the Hugging Face Transformers [14] and Pyserini [15] frameworks, ensuring reproducibility and standardized evaluation. No large language model (LLM) generation was used; all submissions were purely retrieval-oriented.

Sparse retrieval model As a first-stage baseline, we employed the BM25 probabilistic ranking function [3]. BM25 estimates the relevance of a document d to a query q based on term frequency and document length normalization and remains a strong non-neural benchmark for large-scale retrieval tasks. Our implementation used Pyserini's default parameters ($k_1 = 0.9$, $b = 0.4$) for English text.

Neural cross-encoder re-ranking models To improve ranking quality beyond lexical matching, we adopted several transformer-based cross-encoders that operate exclusively as second-stage re-ranking models. These models jointly encode query–passage pairs and directly produce scalar relevance scores, and are not used as standalone dense retrievers.

- **X-Encoder** [4]: a transformer-based cross-encoder that computes relevance scores for query–passage pairs via a shared encoder.
- **MiniLM-6** and **MiniLM-12** [16]: lightweight student models distilled from BERT, balancing ranking effectiveness and computational efficiency. The 6-layer variant prioritizes latency, while the 12-layer variant provides increased contextual capacity.
- **TinyBERT** [17]: a compact, knowledge-distilled transformer model designed to retain BERT-level performance with reduced computational cost. Both global and locally fine-tuned variants were evaluated.
- **ELECTRA** [18]: a transformer encoder pre-trained using a discriminative objective, which has been shown to be effective for relevance estimation tasks.

Re-ranked lists produced by the cross-encoders were truncated to the top 1000 passages per topic and subsequently used for fusion and evaluation.

Fusion methods We also explored hybrid fusion strategies that combine sparse lexical retrieval with neural cross-encoder re-ranking models. Fusion was applied selectively to specific runs, while single-model configurations were evaluated without aggregation.

- **Linear Fusion:** normalized score averaging between BM25 and selected cross-encoder re-ranking models (MiniLM, TinyBERT, and ELECTRA). This approach mitigates model-specific bias and improves robustness across heterogeneous query types.
- **Reciprocal Rank Fusion (RRF)** [5]: an established rank aggregation method that combines independently produced ranked lists. The RRF score for a document d is computed as:

  $$\mathrm{RRF}(d) = \sum_{i=1}^{n} \frac{1}{k + r_i(d)}$$

  where $r_i(d)$ denotes the rank of document d in the ranked list of model i, and k is a smoothing constant (set to 60 in our implementation). RRF is particularly robust to score scale differences and has been shown to improve recall and ranking stability across diverse retrieval configurations.

Implementation and indexing All experiments were conducted on the RAGTIME English subset without translation or multilingual expansion. BM25 indexing and retrieval were performed using the Anserini/Pyserini framework, while neural models were applied exclusively in a second-stage re-ranking setting over the retrieved candidate passages. All ranked lists were stored in standard TREC format for submission and subsequent evaluation by NIST.

3.2.1 Methodology

This section outlines the methodological framework adopted for our participation in the TREC 2025 RAGTIME Multilingual Information Retrieval (MLIR) task.

Our approach focuses on English-only retrieval within the multilingual evaluation setting, emphasizing robustness, reproducibility, and model complementarity. The overall workflow includes data preprocessing, sparse and dense retrieval modeling, hybrid rank fusion, and final evaluation.

3.2.2 Preprocessing and Data Preparation

We utilized the official RAGTIME1 English subset [11], consisting of query–document pairs for ad hoc retrieval. Each topic was preprocessed using standard tokenization and lowercasing. Stopwords were preserved to maintain consistency with dense embedding models. No machine translation or multilingual expansion was applied, ensuring that results reflect pure retrieval performance rather than cross-lingual transfer.

3.2.3 Sparse Retrieval and Neural Re-ranking Pipelines

This work adopts a two-stage retrieval pipeline consisting of sparse lexical retrieval followed by neural cross-encoder re-ranking. In the first stage, BM25 is used to retrieve an initial pool of candidate passages for each topic. In the second stage, transformer-based cross-encoders jointly encode query–passage pairs and assign relevance scores, which are used to re-rank the candidate set. Neural models are applied exclusively in the re-ranking stage and are not used as standalone dense retrievers.

3.2.4 Hybrid Rank Fusion

This subsection describes the fusion strategies applied to selected runs. No fusion is applied to single-model configurations.

3.2.5 Indexing and Retrieval Setup

BM25 indexing and first-stage retrieval were performed using the Anserini/Pyserini framework. Neural cross-encoder models were applied only in a second-stage re-ranking setting over the retrieved candidate passages. No dense vector indexing or approximate nearest neighbor search was employed. All ranked lists were stored in standard TREC format for official evaluation.

3.2.6 Pipeline Summary

In summary, the proposed system follows a lightweight and reproducible two-stage retrieval architecture, combining sparse lexical retrieval with neural cross-encoder re-ranking and optional late fusion. This design emphasizes efficiency, interpretability, and robustness within the MLIR evaluation setting.

3.3 Evaluation Measures

The performance of all retrieval runs was assessed using the official TREC RAGTIME evaluation framework provided by NIST. We adopted standard information retrieval measures designed to capture both precision-oriented and recall-oriented effectiveness. Each system produced a ranked list of up to 1000 documents per topic in TREC format, which was evaluated against relevance judgments using `trec_eval` and the official scoring scripts.

3.3.1 Normalized Discounted Cumulative Gain (nDCG@20)

The primary evaluation measure for the MLIR task was the normalized discounted cumulative gain at rank 20 (*nDCG@20*) [19]. This measure rewards highly ranked relevant documents while penalizing those appearing lower in the ranking. For a given query q, the discounted cumulative gain (DCG) is computed as:

$$\text{DCG@20}(q) = \sum_{i=1}^{20} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

where rel_i denotes the graded relevance of the document at rank i. The normalized form divides DCG by the ideal DCG (IDCG), yielding:

$$\text{nDCG@20}(q) = \frac{\text{DCG@20}(q)}{\text{IDCG@20}(q)}.$$

Final scores are averaged across all queries, providing a balanced measure of early ranking quality that is sensitive to both relevance and rank position.

3.3.2 Mean Average Precision (MAP)

Mean Average Precision (MAP) [20] evaluates overall ranking precision across recall levels. For a given query q, the Average Precision (AP) is defined as:

$$\text{AP}(q) = \frac{1}{R_q} \sum_{k=1}^{N} P@k \cdot rel_k,$$

where R_q is the total number of relevant documents for query q, $P@k$ denotes precision at rank k, and rel_k is an indicator function that equals 1 if the document at rank k is relevant and 0 otherwise. MAP is obtained by averaging AP over all topics, yielding a comprehensive measure of precision across the entire ranking.

3.3.3 Recall@1000

To assess system coverage, we report *Recall@1000*, which measures the fraction of all relevant documents retrieved within the top 1000 ranks. It is formally defined as:

$$\text{Recall@1000}(q) = \frac{|R_q \cap A_{1000}|}{|R_q|},$$

where A_{1000} denotes the set of documents retrieved in the top 1000 positions. This measure emphasizes retrieval completeness and complements the precision-focused measures above.

3.3.4 Evaluation Protocol

Following NIST's official MLIR evaluation guidelines [21], all measures were computed using the TREC RAGTIME 2025 evaluation infrastructure to ensure direct comparability across participants. Runs were validated using `trec_eval` v9.0.7, and scores were averaged across all topics in the English subset. No post-hoc normalization or manual adjustments were applied. All reported measures correspond to automatically verified submissions hosted on the official TREC Evalbase portal.

3.3.5 Summary

The joint use of nDCG@20, MAP, and Recall@1000 provides a robust multi-dimensional evaluation of system performance. nDCG@20 captures early ranking effectiveness, MAP reflects precision across the full ranking, and Recall@1000 quantifies retrieval completeness. Together, these measures enable fair, transparent, and interpretable comparison among heterogeneous retrieval systems participating in the MLIR track.

3.4 Implementation Details

All experiments were conducted using a unified and reproducible evaluation pipeline. BM25 retrieval was implemented using Pyserini, while neural cross-encoder models were evaluated using the Hugging Face Transformers framework. Models were applied in inference-only mode without task-specific fine-tuning. Evaluation was performed using the official TREC RAGTIME infrastructure.

4 Results

4.1 Results Analysis (Auto-ARGUE)

Table 1 reports the official evaluation results for the TREC 2025 RAGTIME Auto-ARGUE task for the submitted runs. Performance is measured using macro-averaged F1, nugget coverage, and citation support, capturing overall argument quality, content coverage, and evidence grounding.

The `tlocal` run achieves the strongest overall performance, obtaining the highest F1$_{macro}$ score (0.149) and nugget coverage (0.090), indicating that locally grounded generation strategies are particularly effective for producing coherent and informative arguments under the Auto-ARGUE framework.

Several MiniLM-based English runs (`eng_min6`, `eng_min6loc`, and `xrc_report`) achieve competitive F1$_{macro}$ scores while exhibiting the highest citation support (0.665), suggesting that lightweight transformer-based pipelines are especially effective at maintaining faithful citation grounding, even when overall argumentative richness is moderate.

In contrast, runs based on larger generative models or more aggressive fusion strategies (e.g., `v2_split_qwen` and `v1_qwen`) achieve higher nugget coverage in isolation but substantially lower F1$_{macro}$ and citation support, highlighting a trade-off between content breadth and argumentative coherence.

Finally, purely lexical or weakly supervised baselines, including `pybm25`, `tf`, and MLP-based variants, exhibit limited performance across all metrics, confirming that effective Auto-ARGUE performance requires at least lightweight neural modeling.

Overall, the results indicate that compact, retrieval-aware transformer-based pipelines provide the most reliable balance between argumentative quality, coverage, and citation faithfulness under strict evaluation constraints.

Table 1: Official RAGTIME Auto-ARGUE results for the submitted DUTH runs, evaluated using macro-averaged F1, nugget coverage (NC), and citation support (CS).

Run ID	**F1$_{macro}$**	**NC$_{macro}$**	**CS$_{macro}$**
tlocal	**0.149**	**0.090**	0.546
eng_min6	0.137	0.094	**0.665**
eng_min6loc	0.137	0.094	**0.665**
xrc_report	0.136	0.093	**0.665**
v2_split_qwen	0.097	0.188	0.253
eng_fused	0.030	0.017	0.556
pybm25	0.030	0.017	0.569
min12	0.014	0.007	0.535
v1_qwen	0.011	0.189	0.139
electra	0.004	0.002	0.646
mini-rrf-report	0.002	0.001	0.655
tf	0.001	0.001	0.568
dtrg-mlp-base	0.000	0.000	0.000
mlp-query-expanded	0.000	0.000	0.000

4.2 Results on Multilingual Retrieval

Table 2 reports the official evaluation results for the TREC 2025 RAG Multilingual Retrieval task for the submitted DUTH runs, as released by the TREC Evalbase. Effectiveness is measured using nDCG@20, MAP, P@20, and Recall@1000, capturing early ranking quality, precision, and coverage.

Early ranking quality (nDCG@20). The `mlir-xenc` and `mlir-pybm25` runs achieve the highest nDCG@20 scores (0.094), indicating stronger early ranking effectiveness for multilingual queries. The `mlir-fused` run follows closely, exhibiting comparable early ranking behavior.

Precision-oriented performance (MAP and P@20). The highest MAP score (0.052) is achieved by the `mlir-xenc`, `mlir-pybm25`, and `mlir-fused` runs, reflecting improved overall precision. Precision@20 remains low but consistent across the strongest configurations, a common characteristic of multilingual retrieval benchmarks with strict relevance criteria.

Coverage and robustness (Recall@1000). Recall@1000 is identical across all submitted runs, indicating comparable document coverage and suggesting that observed differences primarily stem from ranking quality rather than retrieval depth.

Cross-metric interpretation. Overall, the results highlight a trade-off between early ranking effectiveness and model complexity. Lightweight lexical and hybrid configurations achieve competitive early precision, while maintaining robust recall under strict efficiency and reproducibility constraints.

Table 2: Official results for the TREC 2025 RAG Multilingual Retrieval task for the submitted DUTH runs, as reported by the TREC Evalbase.

Run ID	nDCG@20	MAP	P@20	Recall@1000
mlir-xenc	**0.094**	**0.052**	0.007	**0.079**
mlir-pybm25	**0.094**	**0.052**	0.007	**0.079**
mlir-fused	0.093	**0.052**	0.007	**0.079**
mlir-tblocal	0.060	0.016	0.008	**0.079**
mlir-mlm6	0.060	0.019	**0.010**	**0.079**
mlir-mlm6loc	0.060	0.019	**0.010**	**0.079**
mlir-tb	0.035	0.010	0.002	**0.079**
mlir-eng-rrf	0.024	0.003	0.000	**0.079**
mlir-elec	0.012	0.001	0.000	**0.079**
mlir-mlm12	0.009	0.001	0.000	**0.079**

5 Conclusion and Future Work

In this study, we presented the DUTH submission to the TREC 2025 RAGTIME Multilingual Information Retrieval (MLIR) Track, focusing on efficient, transparent, and fully reproducible hybrid retrieval pipelines that combine lexical retrieval with neural cross-encoder re-ranking. The submitted systems rely on carefully calibrated lexical models, lightweight transformer-based re-rankers, and simple fusion strategies, achieving competitive performance across early ranking quality and recall-oriented metrics in a multilingual setting.

The results highlight three main insights. First, localized and multilingual-aware retrieval configurations improve early ranking effectiveness without increasing system complexity. Second, simple fusion strategies such as Reciprocal Rank Fusion (RRF) provide robust performance gains across heterogeneous retrieval signals. Third, strong recall can be achieved using lightweight lexical retrieval in combination with neural re-ranking, even under strict efficiency and transparency constraints.

Future work will explore several directions. One avenue is the incorporation of multilingual fine-tuning on aligned corpora (e.g., mMARCO [22]) to explicitly model cross-lingual representations. Another direction involves the integration of retrieval-augmented generation (RAG) pipelines [23] for evidence-grounded multilingual reporting. We also plan to investigate adaptive fusion strategies that dynamically weight retrievers based on language and domain characteristics. Finally, we aim to open-source our retrieval infrastructure and analysis tools to promote reproducibility within the TREC community [24].

Overall, the findings demonstrate that efficient and interpretable hybrid retrieval pipelines, when combined with robust fusion mechanisms, constitute strong and practical baselines for multilingual retrieval evaluation under realistic efficiency constraints.

References

[1] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[2] Bing Zhang et al. Cross-lingual transfer in retrieval models for multilingual search. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[3] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3 (4):333–389, 2009.

[4] Rodrigo Nogueira and Jimmy Lin. From passage ranking to end-to-end question answering with bert. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[5] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

[6] Georgios Arampatzis, Vasileios Perifanis, Symeon Symeonidis, and Avi Arampatzis. Duth at semeval-2023 task 9: An ensemble approach for twitter intimacy analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1225–1230, 2023. URL `https://aclanthology.org/2023.semeval-1.170/`.

[7] Georgios Arampatzis et al. Duth at exist 2025: Multilingual sexism detection with soft labels and transformers. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[8] Georgios Arampatzis and Avi Arampatzis. Duth at clef joker 2025 tasks 2 and 3: Translating puns and proper names with neural approaches. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[9] Georgios Arampatzis and Avi Arampatzis. Duth at clef 2025 simpletext track: Tackling scientific text simplification and hallucination detection. In *CLEF 2025 Working Notes*, 2025. URL `https://ceur-ws.org/Vol-4038/`.

[10] Ellen M. Voorhees et al. The twenty-ninth text retrieval conference (trec 2025) overview. In *Proceedings of the TREC 2025 Conference*, 2025.

[11] Tri Nguyen et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[12] Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 44th European Conference on IR Research (ECIR)*, 2021.

[13] Ellen M. Voorhees and Donna Harman. Overview of the trec 2005 conference. In *Proceedings of TREC 2005*, 2005.

[14] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP: System Demonstrations*, 2020.

[15] Jimmy Lin et al. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *SIGIR*, 2021.

[16] Wenhui Wang et al. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] Xiaoqi Jiao et al. Tinybert: Distilling bert for natural language understanding. In *Findings of EMNLP*, 2020.

[18] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations (ICLR)*, 2020.

[19] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[20] Ellen M. Voorhees and Donna Harman. The trec-8 question answering track evaluation. In *Proceedings of TREC-8*, 1999.

[21] Ellen M. Voorhees et al. The thirty-fourth text retrieval conference (trec 2025) overview. In *Proceedings of TREC 2025*, 2025.

[22] Luiz Bonifacio, Leonardo Ribeiro, Roberto Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*, 2021.

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[24] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Towards reproducibility in recommender-systems research. In *Proceedings of the 2016 International Conference on Research Challenges in Information Science (RCIS)*, 2016.