PIRB: A Comprehensive Benchmark of Polish Dense and Hybrid Text Retrieval Methods

Sławomir Dadas, Michał Perełkiewicz, Rafał Poświata
National Information Processing Institute, Warsaw, Poland
{sdadas, mperelkiewicz, rposwiata}@opi.org.pl

Abstract

We present Polish Information Retrieval Benchmark (PIRB), a comprehensive evaluation framework encompassing 41 text information retrieval tasks for Polish. The benchmark incorporates existing datasets as well as 10 new, previously unpublished datasets covering diverse topics such as medicine, law, business, physics, and linguistics. We conduct an extensive evaluation of over 20 dense and sparse retrieval models, including the baseline models trained by us as well as other available Polish and multilingual methods. Finally, we introduce a three-step process for training highly effective language-specific retrievers, consisting of knowledge distillation, supervised fine-tuning, and building sparse-dense hybrid retrievers using a lightweight rescoring model. In order to validate our approach, we train new text encoders for Polish and compare their results with previously evaluated methods. Our dense models outperform the best solutions available to date, and the use of hybrid methods further improves their performance.

Keywords: information retrieval, dense retrieval, hybrid retrieval, neural text encoders

1. Introduction

Text information retrieval is a process of retrieving relevant documents from a large collection of text data in response to a user's query. It is a fundamental task in natural language processing and plays a crucial role in various applications, including search engines, question-answering systems, and recommendation engines. Despite the fact that research in the field of information retrieval has a long history, we have recently observed increased interest in this topic. This is primarily related to the emergence of large language models (LLMs), whether offered as services like GPT-4 (OpenAI, 2023) or free and publicly available ones like LLaMA (Touvron et al., 2023a,b) or Falcon (Penedo et al., 2023). With the popularization of these solutions, more attention is also being given to retrieval augmented generation systems (Lewis et al., 2020; Cai et al., 2022), in which a large language model, along with a user's query, receives additional context from an external knowledge base. This context is created based on documents most relevant to the query, extracted using a retrieval algorithm. Correctly selected documents reduce the hallucinations of the language model (Shuster et al., 2021) and allow the injection of additional knowledge that the model may not possess. The quality of the response of such a system is therefore highly dependent on the performance of its retrieval component.

2. Related work

In recent years, an active area of research has been the use of neural networks, particularly Transformer-based language models, to develop novel text retrieval methods. Many new models have been proposed (Yates et al., 2021; Zhao et al., 2022; Guo et al., 2022; Fan et al., 2022), leading to significantly improved results on public benchmarks (Thakur et al., 2021) compared to classical term-based approaches such as BM25 (Robertson et al., 2009). However, this progress has been particularly noticeable for high-resource languages such as English and Chinese, while interest in multilingual and low-resource language solutions has been considerably less pronounced. This can be attributed to the limited availability of datasets, making it challenging to apply the same supervised training methods. This is also related to the lack of standardized benchmarks that would allow for the evaluation of model performance across a wide range of retrieval tasks.

The situation for multilingual retrieval has only recently begun to change with the release of datasets such as mMARCO (Bonifacio et al., 2021), Mr.TyDi (Zhang et al., 2021) and MIRACL (Zhang et al., 2023), each covering more than 10 languages. In the case of models, multilingual encoders designed for semantic textual similarity (Reimers and Gurevych, 2020; Yang et al., 2020) or for bitext mining (Artetxe and Schwenk, 2019; Feng et al., 2022) had been available for several years, but there was a lack of high-quality models adapted for text retrieval problems. This improved with the release of the multilingual versions of the E5 (Wang et al., 2022) models, which were trained in a weakly supervised manner on a large-scale corpus of text pairs extracted from various online data sources, and subsequently fine-tuned on several manually annotated corpora, including

LREC-COLING 2024, pages 12761–12774
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

retrieval data.

Despite these efforts, there are still languages not covered in multilingual text retrieval research. One such language is Polish, which has not been featured in any of the aforementioned datasets. To address this issue, the Polish research community has taken the initiative to prepare language-specific corpora. Rybak et al. (2022) introduced PolQA, a manually annotated dataset for text retrieval containing 7,000 questions matched with passages from Wikipedia. Additionally, some datasets have been created in an automatic or semi-automatic way. Wojtasik et al. (2023) translated the original BEIR benchmark (Thakur et al., 2021) into Polish using Google Translate and evaluated several baseline retrievers and rerankers. Rybak (2023) made available the MAUPQA collection, comprising almost 400,000 automatically generated, translated, or mined question-answer pairs. Since its original release, the collection has been expanded to include more datasets and currently has over one million text pairs. It is also worth mentioning the PolEval-2022 Passage Retrieval challenge (Kobyliński et al., 2023), which utilized data from the PolQA dataset, and introduced two additional subsets from new domains: legal questions and e-commerce FAQ.

Although the amount of Polish data for text retrieval is currently sufficiently large, at least compared to low-resource languages, the vast majority of available corpora have been generated automatically, primarily through machine translation. Manually annotated data make up only a small percentage and are limited to a few thousand queries in a single dataset at most. Consequently, some of the generated datasets may be noisy and contain errors resulting from incorrect translations of documents. Therefore, we believe it would still be beneficial to create new datasets containing real questions and answers written or labeled by humans.

In contrast to datasets, there have been few Polish text retrieval models released to date. As part of the work on the MAUPQA collection, two dense retrievers trained on its question-passage pairs have been published (Rybak, 2023; Rybak and Ogrodniczuk, 2023). Furthermore, the authors of the Polish BEIR benchmark provided four rerankers trained by them[1].

3. Contributions

The aim of our research is to advance Polish text information retrieval in two areas. Our first contribution is to propose a unified benchmark covering a wide range of multi-domain tasks with different characteristics, enabling a reliable and comprehensive evaluation of existing retrieval methods, particularly their generalization ability and zero-shot performance. The second goal is to train and publish new retrieval models for Polish, and then evaluate their performance on the proposed benchmark. The results of our work are described in the article in the following order:

- We present **Polish Information Retrieval Benchmark (PIRB)** covering 41 text retrieval tasks. The benchmark includes pre-existing dataset collections such as MaupQA, BEIR-PL, and PolEval-2022 Passage Retrieval. We have also prepared 10 new, previously unpublished tasks specifically for the evaluation. Nine of these tasks contain sets of actual questions and answers collected from various Polish websites with diverse topics such as medicine, law, business, physics, or linguistics. The last dataset was generated semi-automatically using GPT-3.5 and includes over 8,000 exam-like questions from over 400 university-level courses.
- We perform an evaluation of more than 20 Polish and multilingual text encoders on the PIRB benchmark. The experiments are performed on already existing dense retrieval models, as well as on strong baseline solutions trained by us, including dense retrievers created using scripts from Sentence-Transformers library and the state-of-the-art sparse retrieval model SPLADE (Formal et al., 2021b,a).
- In the last part of the publication, we present our recipe for training highly effective language-specific retrievers. It is a three-step process. First, we use a multilingual knowledge distillation technique (Reimers and Gurevych, 2020) to transfer knowledge from a high-quality English text encoder to a pre-trained language model for Polish. In the next step, we perform a supervised fine-tuning of the created encoder on the annotated retrieval dataset. The final step is to create a lightweight hybrid retriever, combining the results of the sparse and dense methods using an additional learning-to-rank model. Our proposed technique can be an efficient alternative to solutions that combine the retrieval stage with computationally expensive rerankers. In order to validate the effectiveness of the proposed solution, we apply it to train several Polish text encoders and compare their results with previously evaluated methods on the PIRB benchmark.

Furthermore, we make the developed benchmark publicly available[2], as well as the source code of our experiments[3], and the checkpoints of

[1] https://huggingface.co/clarin-knext

[2] huggingface.co/spaces/sdadas/pirb

[3] github.com/sdadas/pirb

all the models we trained[4].

4. Overview of the datasets

In this section, we present the datasets comprising the PIRB benchmark. We begin with a description of the already existing datasets and dataset collections, and then move on to introduce the corpora we have prepared: web datasets and GPT-exams.

4.1. Pre-existing datasets

Of all the datasets included in the PIRB, 31 were previously available as separate benchmarks or individual corpora. When building the benchmark, we sought to collect most of the known publicly available question answering and text retrieval corpora for Polish. We excluded datasets that overlapped with others already found in the benchmark, such as PolQA, whose data was mostly included in queries and passages published for the PolEval-2022 challenge. The following datasets were included in our evaluation:

• **PolEval-2022 Passage Retrieval** was a competition that took place from mid-2022 to early 2023, and its results were summarized during a workshop organized as part of the FedCSIS conference (Kobyliński et al., 2023). During the competition, successive pieces of data were gradually made available, starting from the training (train) and validation (dev-0) splits, which included only data from Wikipedia, to the test sets (test-A and test-B), each containing three subsets from different domains: Wikipedia, legal questions, and e-commerce FAQ. As part of PIRB, we added 7 data subsets from PolEval: dev-0, as well as three parts each for test-A and test-B.

• **BEIR-PL** (Wojtasik et al., 2023) aimed to replicate the original BEIR (Thakur et al., 2021) benchmark for Polish. The authors used Google Translate to translate the datasets and published 11 of them, which we included in PIRB.

• **MAUPQA** (Rybak, 2023; Rybak and Ogrodniczuk, 2023) is an ongoing project aimed at assembling a large and diverse corpus of questions and passages that can be used to train Polish retrievers, rerankers, or generative question answering models. The data was mostly collected automatically, and the author employed various techniques to build individual subsets, including machine translation, generation using large language models, transcription of game show recordings, and utilizing data available in partially structured sources for question and answer mining. Some parts of the MAUPQA collection are exact copies of datasets already present in BEIR-PL, so not all of them were included in PIRB. We selected 12 datasets that were added to the benchmark.

• **MFAQ** (De Bruyn et al., 2021) is a multilingual dataset of questions and answers extracted from FAQ pages found in Common Crawl. While the original collection covered 21 languages, in our evaluation we used only the Polish corpus, consisting of more than 60,000 text pairs.

4.2. Web datasets

The PIRB benchmark includes nine datasets crawled from Polish websites. Our goal was to enrich the benchmark with real-world data written originally in Polish. To accomplish this, we identified websites with separate Q&A sections, some containing questions and answers written directly by the site's editors, and others allowing content creation by registered users. We selected a diverse set of data sources covering various domains such as medicine, law, business, physics, and linguistics. The resulting datasets consist of natural questions, formulated mostly as complete, grammatically correct sentences, as opposed to search engine-oriented datasets such as MS MARCO, in which queries are short and often consist of concatenated keywords. For each data source, we implemented a separate parser to enable precise extraction of question and answer content. We also conducted data cleaning and anonymization, removing personal and contact information. A summary of statistics regarding the collected datasets is presented in Table 1.

The largest datasets originate from websites such as **abczdrowie** and **specprawnik**, where content is created by the user community. These are platforms that connect specialists in a particular field, namely medicine and law, with users seeking answers to questions in those fields. Anyone can ask a question, and any registered and verified specialist can provide an answer, allowing these platforms to build large databases of expert advice. The content on these platforms is moderated or redacted by site editors to a minimal extent, resulting in lower data quality compared to other websites we considered. These datasets also required intensive cleaning and filtering. In the resulting data, we included only answers longer than 200 characters and filtered out all unanswered questions. It is also worth noting that in the case of these two datasets, the average length of questions and answers is similar, as users often describe their personal situation with details. Questions have a more individualized character compared to other platforms.

Another group consists of websites that have a dedicated Q&A section managed by the site staff. Examples of such sites include **e-prawnik**, **gem-**

[4] share.opi.org.pl/s/iS3ziwW9syHdrBj

Dataset name	Domain	Total queries	Avg. words per query	Avg. answers per query	Total documents	Avg. words per document
abczdrowie	medicine	224,533	82	1.22	274,687	77
e-prawnik	law	35,994	63	1.00	35,994	207
gemini	medicine	272	19	1.00	272	120
odi	bussiness	969	8	1.60	1,546	96
onet	trivia	10,943	10	1.00	10,943	27
pwn	linguistics	16,197	41	1.00	16,197	89
specprawnik	law	54,783	66	1.22	66,832	64
techpedia	trivia	9,205	8	1.00	9,205	43
zapytajfizyka	physics	1,446	50	1.00	1,446	211

Table 1: Characteristics of the web datasets collected by us. For each dataset, we report its domain, number of queries and documents, average query and document length in words, as well as the average number of relevant documents per query.

ini, **odi**, **pwn**, and **zapytajfizyka**. Some of these sites allow users to submit their own questions or answers, but these contributions go through editors who review them and decide on their publication. These websites often employ or collaborate with specialists in their respective fields. Datasets created based on these sources are of high quality, but their size is smaller compared to sites relying on user-generated content. Questions are typically shorter and have a general nature, while answers are comprehensive articles providing exhaustive descriptions of the given topic.

The last type of websites includes **onet** and **techpedia**. The datasets based on them come from quizzes or collections of trivia questions, the main purpose of which is not to provide advice in a specific field, but primarily to entertain. Users of these sites can complete quizzes to test their knowledge on various topics. Both questions and answers are short. We discarded most data samples from these sources, as they included answers in the form of a single phrase or a short sentence. Only those cases that had a descriptive answer longer than 50 characters were added to the dataset.

4.3. GPT-exams

GPT-exams is a dataset created by us in a semi-automatic way utilizing the `gpt-3.5-turbo-0613` model available in the OpenAI API. The dataset contains 8,131 exam-like question-answer pairs covering a wide range of topics. To build the dataset, we performed the following steps:

1. We manually prepared a list of 409 university-level courses from various fields. For each course, we instructed the model with the prompt: ”Wygeneruj 20 przykładowych pytań na egzamin z [nazwa przedmiotu]” (Generate 20 sample questions for the [course name] exam). We then parsed the outputs of the model to extract individual questions and performed their deduplication.

2. In the next step, we requested the model to generate the answer to each of the collected questions. We used the following prompt: ”Odpowiedz na następujące pytanie z dziedziny [nazwa przedmiotu]: [treść pytania]” (Answer the following question from [course name]: [question content]). We sent the following system message with the prompt: ”Jesteś ekspertem w dziedzinie [nazwa przedmiotu]. Udzielasz specjalistycznych i wyczerpujących odpowiedzi na pytania.” (You are an expert in [course name]. You provide knowledgeable and comprehensive answers to questions).

3. In the last step, we manually removed from the dataset the cases in which the model refused to answer the question. We searched for phrases such as ”model języka” (language model), ”nie jestem” (I'm not), or ”nie mogę” (I can't). However, such cases were rare, we found less than 10 refusals for the entire dataset.

5. Evaluation

This section provides a description of the evaluation we conducted on the Polish Information Retrieval Benchmark (PIRB). The experiments include both strong baseline models that we trained as a part of this research and other dense text encoders available for the Polish language. First, we describe the evaluated methods, and then we proceed to discuss the obtained results.

5.1. Baseline methods

In our experiments, we included two baseline methods relying on sparse term-based vectors, as well as dense retrievers fine-tuned from Transformer language models. We used the training split of the Polish MS MARCO dataset for all models that required training. The evaluation includes the following sparse approaches:

Model name	Average NDCG@10 (41 tasks)	Avg. without MAUPQA (29 tasks)	PolEval-2022 (7 tasks)	Web Datasets (9 tasks)	BEIR-PL (11 tasks)	MAUPQA (12 tasks)	Other (2 tasks)
Our sparse baselines							
BM25	41.85	43.17	45.51	47.27	33.26	38.64	71.09
SPLADE++	52.93	54.06	58.92	58.60	42.47	**50.22**	80.39
Our dense baselines							
MSE baseline (base)	45.47	47.82	51.87	55.56	33.55	39.80	77.35
MSE baseline (large)	49.98	52.47	57.49	61.13	37.26	43.98	79.58
MNR baseline (base)	46.44	48.98	51.69	55.71	36.52	40.30	77.74
MNR baseline (large)	48.63	51.18	54.22	57.82	38.61	42.45	79.84
Other Polish and multilingual retrievers							
poleval-2022 (base)	45.57	47.62	50.45	55.96	33.84	40.63	75.98
poleval-2022 (large)	48.26	50.90	53.78	58.99	37.41	41.90	78.58
silver-retriever-v1 (base)	$*$	53.61	**60.87**	**61.92**	37.18	$*$	**81.18**
multilingual-e5 (small)	50.65	52.26	57.84	53.82	42.45	46.77	79.72
multilingual-e5 (base)	**53.12**	**55.08**	60.16	59.09	**44.01**	48.38	80.18
multilingual-e5 (large)	**57.29**	**60.09**	**65.86**	**64.35**	**48.99**	**50.53**	**81.81**

Table 2: Evaluation results of our baselines and other available retrieval models for Polish. We report Normalized Discounted Cumulative Gain at 10 (NDCG@10) for the entire PIRB benchmark consisting of 41 tasks as well as for separate task groups. The best score in each column is shown in blue, the second best in red.
$*$ Since silver-retriever-base-v1 (Rybak and Ogrodniczuk, 2023) was trained on datasets from MAUPQA, it cannot be reliably evaluated on this part of the benchmark. In order to compare it to the other models, we also provide average scores on the benchmark with MAUPQA datasets excluded.

- **BM25** (Robertson et al., 2009) is a popular and effective term-based ranking function used in information retrieval since the 1980s. It is an extension of TF-IDF weighting scheme, providing a balanced approach to text ranking, considering both term frequency and document specificity. In our experiments, we used the implementation of BM25 available in anserini (Yang et al., 2018). We applied the default Polish analyzer, which performs the lemmatization of words with the Morfologik[5] library.
- **Sparse Lexical and Expansion (SPLADE)** model (Formal et al., 2021b,a) is a family of modern term-based methods employing Transformer language models. In this approach, the masked language modeling (MLM) head is optimized to generate a vocabulary-sized weight vector adapted for text retrieval. SPLADE is a highly effective sparse retrieval ranking algorithm, achieving better performance than classic methods such as BM25. Unlike those methods, SPLADE directly uses subwords generated by the tokenizer as terms. We trained the Polish version following the same procedure as the **SPLADE++** variant (Formal et al., 2022), utilizing hard negatives mined with an ensemble of cross-encoders (EnsembleDistil). We used Polish DistilRoBERTa (Dadas et al., 2020) as our base language model.

In addition to the above methods, we also trained Polish text encoders using scripts provided in the Sentence-Tranformers library. These examples show how to train high-quality dense retrievers with a set of hard negatives acquired from a cross-encoder. Authors of the library propose two fine-tuning methods with different loss functions:

- **MSE (Mean Squared Error) baseline** is a model fine-tuned using the Margin-MSE loss (Hofstätter et al., 2020), which aims to reduce the margin difference for a positive-negative pair between the reference scores obtained from cross-encoder and scores produced by the optimized model. Specifically, the loss is calculated using the following formula:

$$\text{MSE}\Big(s(q,d^+) - s(q,d^-), \hat{s}(q,d^+) - \hat{s}(q,d^-)\Big) \quad (1)$$

in which q, d^+ and d^- denote query, positive document, and negative document, respectively, s is the reference similarity between query-document pair, and $\hat{s}$ is the similarity computed by the trained model.

- **MNR (Multiple Negatives Ranking) baseline** uses contrastive loss with in-batch negatives (Henderson et al., 2017) in addition to hard negatives. Given a query q_i and a mini-batch consisting of K queries, the loss for that query is calculated with

[5] https://github.com/morfologik/morfologik-stemming

the following formula:

$$-\hat{s}(q_i, d_i^+) + \log \sum_{\substack{k=1,\\ k\neq i}}^{K} e^{\hat{s}(q_i, d_k^+)} + \log \sum_{k=1}^{K} e^{\hat{s}(q_i, d_k^-)} \quad (2)$$

in which d_k^+ and d_k^- denote positive and negative document for the k-th query, and $\hat{s}$ is the computed text pair similarity.

We fine-tuned Polish RoBERTa base and large (Dadas et al., 2020) language models with both methods. We used the default hyperparameters defined in the original training scripts, except for the batch size, which we reduced from 64 to 32 for the large models. The models were trained for 10 epochs. We employed a learning rate scheduler with a warmup phase for the first 1,000 batches, a peak learning rate of $2e{-}5$, and linear decay for the remainder of the training. Cosine similarity was used as a similarity metric and mean pooling as the output pooling method.

5.2. Other methods

As part of our experiments, we evaluated over 20 publicly available Polish and multilingual dense text encoders, but only a few of them proved to be effective enough for practical use in retrieval tasks. In this publication, we present results only for the models that achieved an NDCG@10 score higher than the BM25 baseline. In the full ranking, which we have made available online[6], we include all of the evaluated retrievers, along with more detailed results covering individual datasets. Models offering good performance for Polish text retrieval include:

• **Multilingual E5** is a text encoder supporting over 100 languages. It is a multilingual version of the E5 model (Wang et al., 2022), developed using the same two-stage training procedure. The first stage involved weakly-supervised training on a dataset of text pairs extracted from large internet corpora, such as Common Crawl. In the second stage, the model was fine-tuned in a supervised manner on several annotated datasets. The multilingual versions, which were released a few months after the original English models, were made available in small (118M parameters), base (278M), and large (560M) sizes.

• **Silver Retriever** (Rybak and Ogrodniczuk, 2023) is a Polish dense retrieval model trained on MAUPQA datasets with hard negatives mined employing a combination of heuristic rules and cross-encoders. The model was trained with constrastive loss for 15,000 steps using a batch size of 1024. Only the base-sized version of the encoder was released, fine-tuned from the HerBERT language model (Mroczkowski et al., 2021).

• Although the solutions submitted to the **PolEval-2022** Passage Retrieval competition focused primarily on the reranking phase, a few dense retrievers were also trained. Kozłowski (2023) provided two text retrieval models, base and large, fine-tuned on a Polish translation of MS MARCO and training data from the competition, with a set of negatives generated using BM25.

5.3. Results

The results of our evaluation are presented in Table 1. The table compares the performance of our baseline models and the other models described in the previous section. The methods are scored according to the Normalized Discounted Cumulative Gain at 10 (NDCG@10) metric.

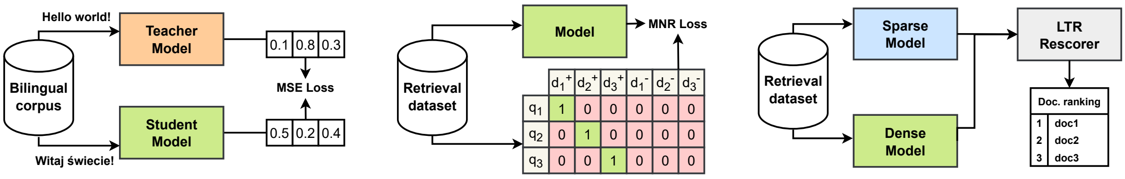
We can see that the multilingual E5 models demonstrate high quality, with the large model significantly outperforming the other evaluated solutions. Among the smaller models, in addition to E5 base, Silver Retriever and SPLADE++ also deliver good results. Silver Retriever achieves the second-highest score in three task groups, although worse performance on the BEIR-PL datasets lowers its overall rating. SPLADE++ performs well on all task groups, achieving an average NDCG@10 value close to E5 base, despite being based on a model with only 82 million parameters, fewer than any of the evaluated dense retrievers.

Models from the PolEval-2022 competition and our dense baselines lag behind the other methods. All of these retrievers show similar performance, with MSE large baseline being marginally better than the other models. However, the difference in NDCG@10 between these models and E5 with a similar number of parameters is at least a few points in favor of E5.

6. Dense and hybrid retrievers

The final stage of our research involved training new dense text encoders for the Polish language that could compete with the best currently available models. In this section, we present our recipe for training effective language-specific retrieval methods. Our proposed process consists of three steps, as illustrated in Figure 1. We begin this section by providing an overview of each stage of building our retrieval solution. We then compare the results achieved by our models on the PIRB benchmark with methods tested earlier.

[6] huggingface.co/spaces/sdadas/pirb



① Distillation from high-quality English teacher model

② Fine-tuning using contrastive loss with in-batch and hard negatives

③ Building a lightweight sparse-dense hybrid with LTR rescorer

Figure 1: An overview of our procedure for building effective retrieval methods. In the first step, we perform knowledge transfer from an English dense retriever using a bilingual corpus. Next, we fine-tune the obtained model on an annotated dataset for text retrieval in the target language using contrastive loss. In the final step, we construct a lightweight hybrid combining dense and sparse methods, utilizing an additional learning-to-rank model.

6.1. Methodology

Below, we present a description of the individual stages of our experiments. The dense retrieval model resulting from each of the steps is used in the subsequent stage of the process. The order of these steps is as follows:

1. **Knowledge distillation** - Our procedure begins with transferring knowledge from a high-quality English text encoder to a language-specific model, employing multilingual knowledge distillation technique (Reimers and Gurevych, 2020). In this procedure, the original well-performing encoder is known as the teacher model, while the new encoder we want to train is the student model. We initialized the student with the weights of an already trained checkpoint supporting the target language, in our case Polish. We selected two groups of models of different sizes as the basis for our experiments: pre-trained Polish RoBERTa language models (Dadas et al., 2020) and multilingual E5 (Wang et al., 2022). For teachers, we chose English FlagEmbedding (Xiao et al., 2023) models. The goal of this knowledge distillation method is to fine-tune a student model using bilingual corpora to approximate text representation generated by the teacher. During training, the original encoder generates a vector representation of the English text, while the student produces a representation for the translation of that text. The difference between the vectors is then used to compute the mean-squared error (MSE) loss, which aims to reduce the distance between these representations.

To conduct this stage of training, we collected a Polish-English bilingual corpus consisting of over 60 million text pairs. The corpus included subsets representing various types of content: 1) single sentences and short texts from the OPUS (Tiedemann, 2012) project; 2) longer paragraph-aligned texts from PELCRA (Przepiórkowski et al., 2010), DGT (Steinberger et al., 2012), Wikipedia, and bilingual abstracts from scientific publications; 3) a collection of over 2 million questions extracted from Common Crawl and question-answering datasets, translated from English to Polish using machine translation. The collected texts were subsequently filtered by the LaBSE (Feng et al., 2022) model to exclude noisy and low-quality translations. We discarded texts for which the similarity between aligned sentences was lower than 0.7. The models were trained with a batch size of 32. We trained small and base models for 3 epochs, large models for 2 epochs. A learning rate scheduler with linear decay and a warmup phase of 10,000 steps was used.

2. **Fine-tuning** - The second stage involved further supervised training on an annotated text retrieval dataset. This step is similar to the previously used procedure for training baseline MNR models. In this case, we also used the training split of the MS MARCO dataset, optimizing the models with contrastive loss utilizing in-batch and hard negatives. However, we applied different training hyperparameters. Following Wang et al. (2022), we introduced a temperature hyperparameter and set it to a value of 0.01. We also used larger batch sizes: 288 for small models, 192 for base models, and 72 for large models. All models were fine-tuned for 50,000 steps, with a warm-up phase for the first 1,000 steps to reach a peak learning rate of $2e-6$, which was then linearly decayed.

3. **Sparse-dense hybrids** - The final step in our procedure was to build a hybrid system that combines results from dense and sparse models. Since these representations can capture different characteristics of the input data, merging results from two indexes can improve the overall performance of the text retrieval system (Chen et al., 2022; Luan et al., 2021). Naive aggregation methods involve computing new weights for query-document pairs as a weighted average of the scores returned by each index. Unfortunately, due to the different scales of the scores, it's difficult to set such weights that would provide good zero-

Student model	Teacher model	Before fine-tuning	Fine-tuned	Hybrid (BM25)	Hybrid (SPLADE)
No distillation, only fine-tuning					
multilingual-e5 (small)	-	50.65	51.38 (+0.73)	53.56 (+2.91)	55.63 (+4.98)
multilingual-e5 (base)	-	53.12	54.13 (+1.01)	55.59 (+2.47)	56.79 (+3.67)
multilingual-e5 (large)	-	**57.29**	57.44 (+0.15)	58.47 (+1.18)	58.66 (+1.37)
Distillation + fine-tuning					
multilingual-e5 (small)	bge-small-en	47.64	52.25 (+4.61)	54.25 (+6.61)	56.04 (+8.40)
multilingual-e5 (base)	bge-base-en	53.56	56.01 (+2.45)	56.99 (+3.43)	57.91 (+4.35)
multilingual-e5 (large)	bge-large-en	**56.00**	**58.03** (+2.03)	**58.84** (+2.84)	**59.22** (+3.22)
polish-roberta-v2 (base)	bge-base-en	53.60	56.11 (+2.51)	57.20 (+3.60)	58.01 (+4.41)
polish-roberta-v2 (large)	bge-large-en	55.62	**58.06** (+2.44)	**58.82** (+3.20)	**59.23** (+3.61)

Table 3: A comparison of the NDCG@10 metric on the PIRB benchmark for distilled, fine-tuned and hybrid methods. The best score in each column is shown in blue, the second best in red.

shot results for different datasets. More complex systems use two-stage retrieval, in which the first stage retrieves a number of relevant documents from each index, and then the documents are sorted by the reranker model. However, such systems are considerably slower than purely retrieval-based solutions since the most common reranker architectures require calculating the score for each query-document pair separately.

In our approach, we propose a lightweight learning-to-rank model that takes the scores of dense and sparse models as input and returns a new score for each query-document pair. To train the model, we use the pairwise LambdaMART (Burges, 2010) algorithm implemented in the XGBoost (Chen and Guestrin, 2016) library, which optimizes the NDCG metric. For each index that makes up the hybrid system, four attributes are passed to the learning-to-rank model: the score for a given query-document pair, the maximum and minimum score returned for the entire list of relevant documents, and a binary attribute indicating whether a given document is on the relevant documents list. If the document is not present in the set of relevant documents, all four attributes for the index are set to 0. As in the previous stage, we also use a training split of Polish MS MARCO to optimize the rescoring model. We use the following hyperparameters for the XGBRanker model: number of gradient boosted trees is set to 100, max tree depth to 6, subsampling ratio to 0.75, and column subsampling by tree to 0.9. In our experiments, we evaluate two types of hybrid indexes: one using the BM25 method as the sparse index and another utilizing the stronger SPLADE model.

One of the advantages of the proposed rescoring model is its efficiency, resulting in minimal computational overhead on the entire retrieval system. For a hybrid system composed of two indexes and a maximum number of relevant documents set to 200 per query (100 from each index), the throughput of aggregating results using our lightweight learning-to-rank model was approximately 1,500 queries per second on a machine with Intel i7-13700 CPU and Nvidia RTX 3090 GPU. For comparison, using a Transformer-based reranker[7] on the same configuration achieved a throughput ranging from 2 to 5 queries per second, depending on the dataset and length of the documents in batch.

6.2. Results

The results of our experiments are presented in Table 3. We compared the average NDCG@10 values achieved by distilled, fine-tuned, and hybrid models on the PIRB benchmark. We also included an evaluation of the original multilingual E5 models, which exhibited high performance in retrieval tasks. We fine-tuned them on Polish data and built sparse-dense hybrids, just like for the distilled models.

We can observe that after the distillation stage, the quality of the obtained models is lower than that of the available E5 models. An exception is the base-sized models, which already gain a slight advantage at this stage. However, distilled models respond significantly better to fine-tuning. Each model achieves better results after that stage, with distilled models experiencing an increase in the NDCG@10 metric ranging from 2.5 to over 4.5 points. For the original E5 models, this observed increase is lower, at a maximum of 1 point.

The use of hybrid solutions allows for further improvement in results. The observed increases are higher for smaller dense models, for which the combination with a sparse index yields strong results. For instance, the hybrid of the small-sized model with the SPLADE model significantly outperforms standalone base-sized models, even approaching results obtained by large models by

[7] https://huggingface.co/clarin-knext/herbert-base-reranker-msmarco

around 1 point. For the other model sizes, an increase is also noticeable, with base models gaining an additional 2 points and large models gaining 1 point in NDCG@10 compared to the results from the second stage. As expected, SPLADE hybrids offer better quality than BM25 hybrids, but the difference between these methods also decreases as the size of the dense model increases.

7. Conclusion

In this work, we have introduced PIRB, a text information retrieval benchmark for the Polish language consisting of 41 tasks, including 10 that incorporate new, previously unpublished datasets. We conducted a detailed evaluation of multiple retrievers, including dense and hybrid solutions trained as part of our research. The methods built using our approach achieved results surpassing the best text retrieval models for Polish to date. We believe that our work will help standardize the evaluation of information retrieval systems and advance research in this area for Polish.

8. Acknowledgements

This research was conducted with the A100 GPU cluster support delivered by the Gdansk University of Technology within the TASK center initiative.

9. Bibliographical References

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.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset.

Christopher JC Burges. 2010. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11(23-581):81.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. 2022. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3417–3419.

Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. 2022. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In *European Conference on Information Retrieval*, pages 95–110. Springer.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing: 19th International Conference, ICAISC 2020, Zakopane, Poland, October 12-14, 2020, Proceedings, Part II 19*, pages 301–314. Springer.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, Jiafeng Guo, et al. 2022. Pre-training methods in information retrieval. *Foundations and Trends® in Information Retrieval*, 16(3):178–317.

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 (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Thibault Formal, C. Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *ArXiv*, abs/2109.10086.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–42.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Łukasz Kobyliński, Maciej Ogrodniczuk, Piotr Rybak, Piotr Przybyła, Piotr Pęzik, Agnieszka Mikołajczyk, Wojciech Janowski, Michał Marcińczuk, and Aleksander Smywiński-Pohl. 2023. PolEval 2022/23 challenge tasks and results. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, page 1237–1244. IEEE.

Marek Kozłowski. 2023. Hybrid retrievers with generative re-rankers. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, page 1265–1270. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. Herbert: Efficiently pretrained transformer-based language model for polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10.

R OpenAI. 2023. GPT-4 technical report. *arXiv*, pages 2303–08774.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Adam Przepiórkowski, Rafał L Górski, Marek Łazinski, and Piotr Pezik. 2010. Recent developments in the national corpus of polish. *NLP, Corpus Linguistics, Corpus Based Grammar Research*, page 302.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Piotr Rybak. 2023. Maupqa: Massive automatically-created polish question answering dataset. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 11–16.

Piotr Rybak and Maciej Ogrodniczuk. 2023. Silver-retriever: Advancing neural passage retrieval for polish question answering.

Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk. 2022. Improving question answering performance through manual annotation: Costs, benefits and strategies. *arXiv preprint arXiv:2212.08897*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the*

Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2214–2218.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Konrad Wojtasik, Vadim Shishkin, Kacper Wołowiec, Arkadiusz Janz, and Maciej Piasecki. 2023. Beir-pl: Zero shot information retrieval benchmark for the polish language. *arXiv preprint arXiv:2305.19840*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding.

Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible ranking baselines using lucene. *J. Data and Information Quality*, 10(4).

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

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. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*.

A. Appendix: Data pre-processing

This section describes our data pre-processing pipeline for Web Datasets. Since they contain questions and answers collected from the internet, additional cleaning and filtering steps are required to ensure the appropriate quality of evaluation data. For the GPT-exams, there is no need to perform the described steps because the model's responses do not exhibit the same issues. The cleaning process primarily relies on heuristics and dictionary-based methods. More specifically, we employed the following steps:

- **Text normalization** - We removed e-mails and www addresses using regular expressions. Additionally, special characters were removed, and sequences of whitespace characters were replaced with a single space.

- **Removing personal information** - In order to anonymize the documents, we split the text into sentences and lemmatized each sentence. Then, we use a set of dictionaries of Polish first names, surnames, and words indicating contact information (such as "mail," "www," "tel", "address") to identify sentences containing such words and remove them from the texts.

- **Removing non-informative phrases** - Using a dictionary-based method, we also removed phrases from the beginning and end of the text, both from questions and answers. We removed texts such as "good morning", "thank you in advance", "best regards", which do not contribute substantively to the content.

- **Numbering removal** - In some sources, questions are numbered. For these datasets, we remove the number if it appears at the beginning of the question.

- **Removing questions with images** - In two datasets (techpedia and onet), there may be questions that contain additional context in the form of an image. In the case of these datasets, we removed all questions containing phrases such as "picture", "photo", "image".

- **Removing short texts** - Most erroneous and noisy samples come from short questions or answers. Moreover, short answers are often too general to be matched to a specific question without providing additional context. Therefore, we removed questions shorter than 10 characters and answers shorter than 50 characters (200 for abczdrowie and specprawnik, since those two datasets were of lower quality than the rest).

The source code of our preprocessing script, along with the dictionaries and execution arguments for each dataset, is available here: https://share.opi.org.pl/s/oMFJqJ5sSWjzFSX.

B. Appendix: Error analysis

In this section, we present conclusions from error analysis based on the results obtained by our standalone retrieval models and hybrid methods. In order to further investigate the results of our methods, we conducted two comparisons. Below is the summary from the analysis.

Best of our dense retrievers (distilled and fine-tuned polish-roberta-large-v2) vs. previous state-of-the-art (multilingual-e5-large):

- We divided all datasets into subgroups based on average query length: short (less than 10 words), medium (10-20 words), and long (more than 20 words). Our model outperforms E5 on 53% of short, 50% of medium, and 83% of long datasets.

- PolEval-2022 is the only taks group, in which E5 performs significantly better. The group consists of seven datasets, all with a similar structure, comprising short natural queries and relatively short passages (below 50 words).

- On the other hand, our model achieves better results on datasets from the BEIR-PL and Web Datasets groups, which consist of a more diverse sets of tasks. BEIR-PL also includes datasets with short queries from search engines, such as MS MARCO, NFCorpus, or DBPedia. On these datasets, our model has a few points of NDCG@10 advantage over E5.

- E5 performs better or comparably to our model on trivia-like datasets, while our model exhibits better generalization ability to specialized domains.

Our standalone retrievers vs. hybrid retrieval:

- The use of a hybrid retrieval improves results primarily for queries containing named entities or specialized terminology. Such cases can benefit from term-based matching of sparse indexes.

- In addition, the use of the hybrid approach improves results for task groups, in which the standalone model performed worse or comparably to the original E5 model, particularly for PolEval-2022 and MAUPQA tasks. In the other groups, the differences are small.

C. Appendix: Detailed results

In the main part of the publication, we focused only on the subset of models evaluated by us, which we assessed according to the NDCG@10 metric. Below are more detailed results of our experiments, demonstrating a wider range of metrics including NNDCG@10, MRR@10, Recall@100, and Accuracy@1.

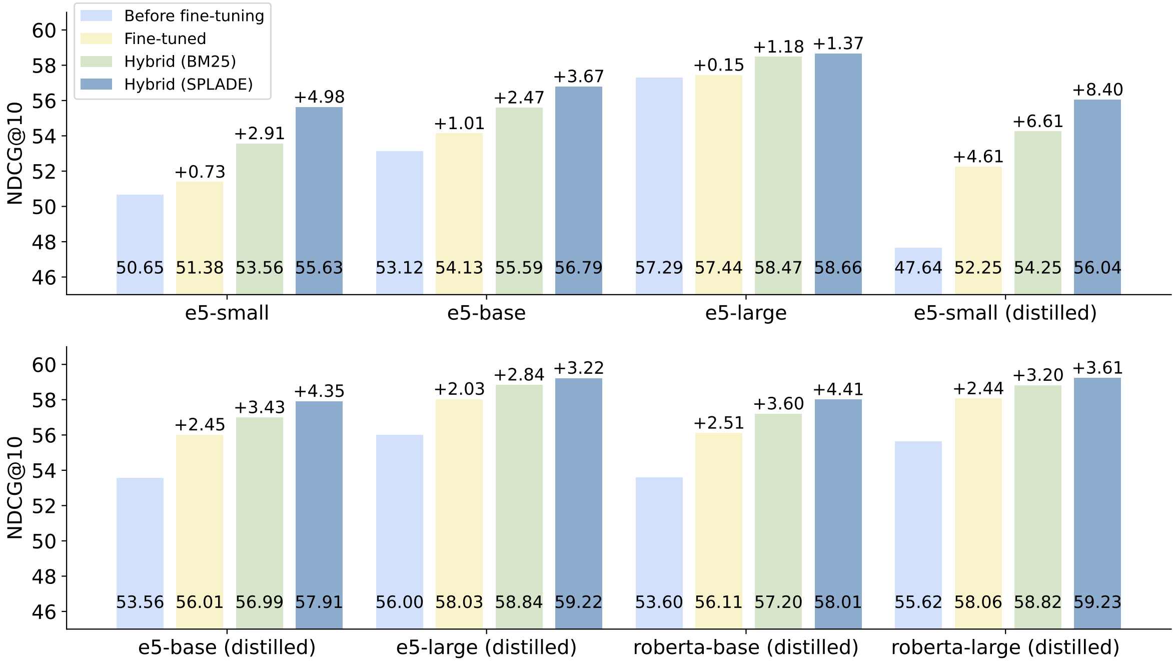


Figure 2: NDCG@10 values obtained by the models trained in our study. We present the results of the original multilingual E5 models, as well as models distilled from FlagEmbeddings based on E5 and Polish RoBERTa. For each model, we show its performance before fine-tuning, after fine-tuning on the Polish MS MARCO dataset, and the result of the sparse-dense hybrid combining the given model with BM25 and SPLADE methods.

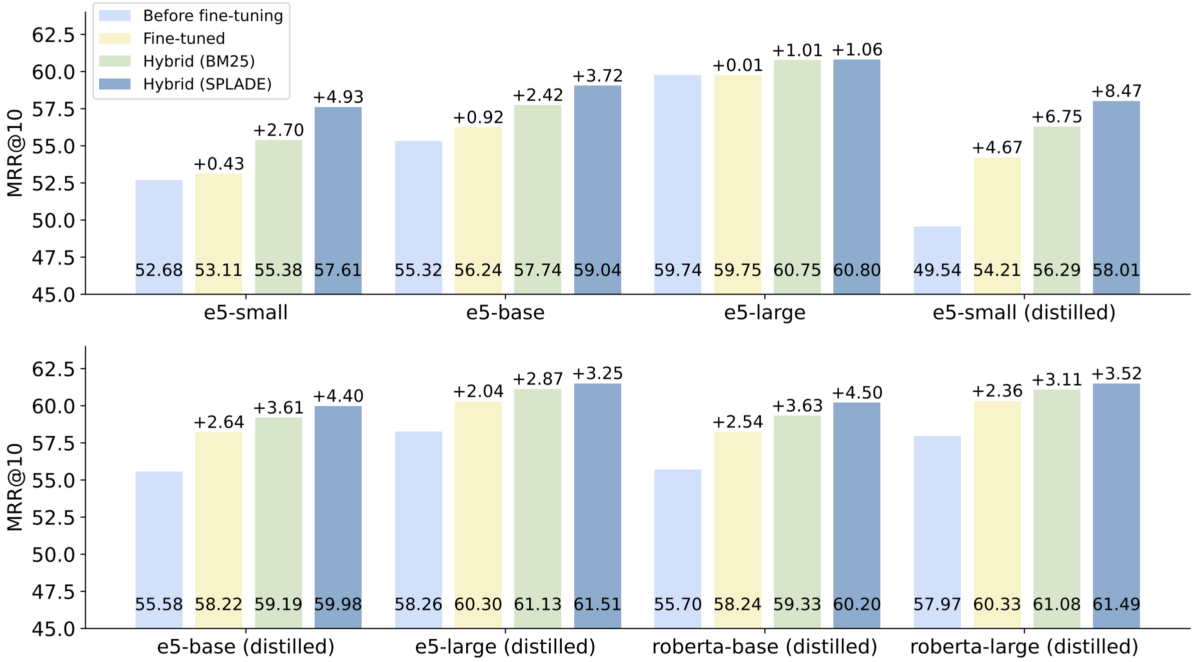


Figure 3: MRR@10 values obtained by the models trained in our study. We present the results of the original multilingual E5 models, as well as models distilled from FlagEmbeddings based on E5 and Polish RoBERTa. For each model, we show its performance before fine-tuning, after fine-tuning on the Polish MS MARCO dataset, and the result of the sparse-dense hybrid combining the given model with BM25 and SPLADE methods.

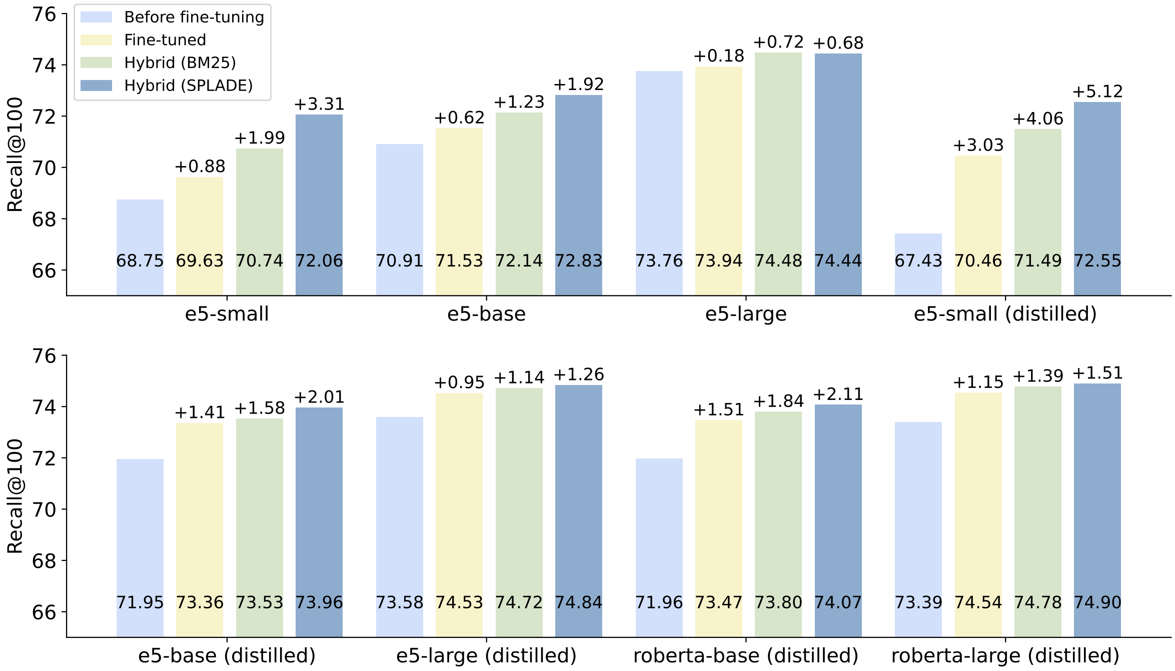


Figure 4: Recall@100 values obtained by the models trained in our study. We present the results of the original multilingual E5 models, as well as models distilled from FlagEmbeddings based on E5 and Polish RoBERTa. For each model, we show its performance before fine-tuning, after fine-tuning on the Polish MS MARCO dataset, and the result of the sparse-dense hybrid combining the given model with BM25 and SPLADE methods.

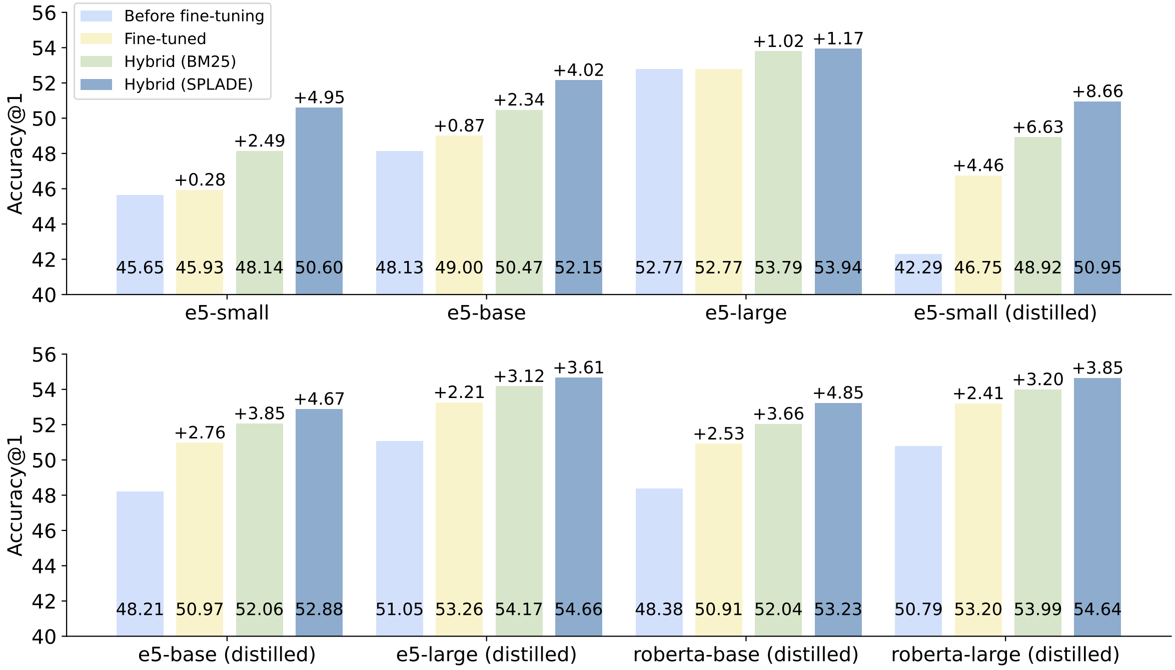


Figure 5: Accuracy@1 values obtained by the models trained in our study. We present the results of the original multilingual E5 models, as well as models distilled from FlagEmbeddings based on E5 and Polish RoBERTa. For each model, we show its performance before fine-tuning, after fine-tuning on the Polish MS MARCO dataset, and the result of the sparse-dense hybrid combining the given model with BM25 and SPLADE methods.