Team Galápagos Tortoise at LongEval 2024

Neural Re-Ranking and Rank Fusion for Temporal Stability

Marlene Gründel[1,†], Malte Weber[1,†], Johannes Franke[1,†] and Jan Heinrich Reimer[1]

[1] *Friedrich-Schiller-Universität Jena, 07743 Jena, Germany*

Abstract

We describe the participation of team Galápagos Tortoise in the LongEval shared task at CLEF 2024. We aim to construct a highly effective retrieval system that, unlike many popular modern models, retains its effectiveness over a long period of time. To this extent, we follow two approaches: First, we experiment with different schemes to aggregate passage scores of monoT5 re-rankings. Second, we propose a weighted rank fusion of retrieval models implementing different paradigms: RankZephyr, a sparse cross-encoder, ColBERT, and BM25. Our key findings indicate that, despite our efforts, all systems exhibit a temporal decline in effectiveness. While using monoT5 with max passage aggregation outperforms mean passage aggregation on all datasets—over longer periods even more significantly—we find that monoT5 is generally too sensitive towards long-term changes to observe meaningful differences when using another aggregation scheme. Moreover, our rank fusion approach, although dominated by RankZephyr, achieves higher effectiveness than the individual fused models but is also more prone to long-term instability. This emphasizes the importance of developing hybrid models combining lexical and neural systems to obtain highly effective retrieval systems but also shows that to achieve sustainable effectiveness, the fusion components must be selected carefully.

Keywords

Longitudinal evaluation, neural ranking, rank fusion

1. Introduction

Modern retrieval systems typically use a multi-stage re-ranking architecture, where the results of a recall-oriented (typically lexical) first-stage ranker are subsequently refined with precision-oriented (typically neural) re-rankers [1, 2, 3, 4, 5, 6]. Such multi-stage models perform well on test collections like MS MARCO [7, 8]. Static ad-hoc test collections, however, are prone to train-test leakage [9, 10, 11] and do not resemble the realistic scenario where documents and the use of language change over time or new documents become available [12, 13]. Current state-of-the-art models are typically trained on a fixed dataset containing only documents up to a specific point in time [14]. These models trained on fixed-in-time data struggle to maintain their effectiveness when applied to more recent datasets [15, 16, 17].

The LongEval lab explores the extent to which temporal declines in retrieval effectiveness occur with different retrieval paradigms and aims to support the development of retrieval systems that are persistent in their effectiveness over time [18, 19]. Systems are evaluated on three test sets of three months of documents and queries from query logs of the French web search engine Qwant[1] in 2023.

We experiment with combining more stable lexical with less stable but highly effective neural retrieval systems in order to develop effective and long-term stable systems. We evaluate two distinct approaches in our submissions to the LongEval shared task: (1) For the popular monoT5 [8] cross-encoder model, we evaluate the effect of using more than the best-scoring passage when aggregating the document score after a lexical first-stage retrieval. Our assumption is that a good trade-off between effectiveness and temporal robustness can be achieved when averaging the scores from the top-k documents for an optimal k. And (2) we test if a rank fusion of a variety of effective lexical and neural retrieval systems

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9–12, 2024, Grenoble, France

[†] These authors contributed equally.

marlene.gruendel@uni-jena.de (M. Gründel); malte.weber@uni-jena.de (M. Weber); johannes.franke@uni-jena.de (J. Franke); heinrich.reimer@uni-jena.de (J. H. Reimer)

0000-0003-3470-1326 (M. Gründel); 0009-0006-7849-7864 (M. Weber); 0009-0007-8788-2374 (J. Franke); 0000-0003-1992-8696 (J. H. Reimer)

[1] https://qwant.com/

is more robust to temporal changes than a single state-of-the-art re-ranking model based on a large language model (LLM). By using multiple systems trained on different datasets or completely unaware of training data, we seek to improve long-term stability while not degrading effectiveness.

To this extent, we submit five runs to the LongEval shared task and test hypotheses grounded on the aforementioned assumptions. Three runs use a combination of BM25 and PL2 lexical retrieval with Bo1 query expansion and monoT5 re-ranking, then aggregating monoT5's passage-level scores with different aggregation schemes by averaging the score of a subset of the passages. The remaining two runs are our proposed weighted rank fusion of RankZephyr, a sparse cross-encoder model, ColBERT, and BM25, as well as just RankZephyr as a baseline.[2]

The results for our first group of runs using a combination of lexical first-stage retrieval, query expansion, and monoT5 re-ranking show (1) that the nDCG effectiveness of monoT5 re-ranking still declines over time when using top-4 average passage aggregation, (2) that the choice of the passage aggregation scheme does only marginally impact the overall effectiveness, but also (3) that the difference in nDCG between the aggregation schemes gets more pronounced over time. Our rank fusion of RankZephyr with neural and lexical models slightly improves the effectiveness. Yet, both the rank fusion and RankZephyr demonstrate stronger long-term instability than the other methods examined. Combinations of lexical and neural systems can therefore increase the effectiveness of retrieval systems, but are not necessarily accompanied by increased stability. Further research is needed to identify fusion components that achieve sustainable effectiveness.

2. Related Work

Our submission builds on prior work that proposed a way to use monoT5 in a multi-stage document re-ranking, utilizing document expansion (e.g., using T5 [20]) to enrich documents with their keyword representation [8]. In our approach, we also perform a query expansion, although with Bo1 [21] instead, and we use PL2 [22] and a BM25 [23] scoring as first-stage retrieval. In particular BM25 is applied in many multi-stage re-ranking architectures to retrieve candidate documents for subsequent re-ranking [3]. As these multi-stage re-rankers are often limited by the used models' context window, usually, after retrieval documents are split into shorter text passages which are passed to the re-rankers [24]. After re-ranking, several strategies are applied to aggregate the passage-level scores [25]

As our second line of research, rank fusion combines rankings returned by multiple search engines such that the combination maximizes a certain effectiveness criterion. Previous works have shown that such combinations consistently improve retrieval effectiveness [26, 27, 28, 29, 30]. In our work, we fuse four different retrieval models: BM25 [23], a sparse cross-encoder [31], ColBERT [32] and RankZephyr [33]. We employ BM25 for its robustness and frequent use in similar research [3]. Cross-encoders are effective [34, 6, 35] but often inefficient with respect to their inference run time, memory footprint, and energy consumption [36]. Compared to full attention as used in monoT5, Schlatt et al. [31] improved the efficiency while maintaining effectiveness by combining windowed self-attention and asymmetric cross-attention between sub-sequences [31]. We use their efficient yet effective cross-encoder model as another model for our rank fusion approach. ColBERT [32] is also used in our rank fusion due to implementing a completely different retrieval paradigm, late interaction. With late interaction, ColBERT strives to reconcile efficiency and contextualization while estimating the relevance of a document for a given query. Finally, we integrate RankZephyr [33], an open-source LLM for listwise zero-shot re-ranking, that outperforms GPT-4 [37] effectiveness on several datasets.

In our system implementations, we use ranx.fuse [38], a Python library for rank fusion and PyTerrier [39]. The PyTerrier framework implements a wide range of lexical first-stage retrieval models, such as PL2 [22] and a BM25 [23], and also allows for composing multi-stage retrieval pipelines [40]. The LongEval datasets were accessed via ir_datasets [41] and its TIREx integration [42] which allowed us to use the same containerized software during development and submission, and to archive the submission code on TIRA [43].

[2] Code and data available online: https://github.com/tira-io/ir-lab-jena-leipzig-wise-2023-galapagos-tortoise/

3. Approach

With our participation in the LongEval shared task, we pursue two different ranking approaches: First, we compare retrieval pipelines that implement neural re-ranking with monoT5 but use differing passage score aggregations. Further, we tune a weighted rank fusion of RankZephyr, a sparse cross-encoder, ColBERT, and BM25 towards maximizing nDCG@10 on the LongEval data collection from January 2023 [44].

3.1. Neural Re-Ranking with monoT5

Our initial retrieval pipeline consists of a weighted linear score combination of a PL2 scoring [22] and a BM25 scoring [23] with Bo1 query expansion [21], the latter (BM25+Bo1) being weighted twice as high as the PL2 score. The motivation behind our choice for this initial retrieval stage is to aim for increasing temporal stability with a fused system of two lexical approaches, but at the same time, not to tune the weights on the training data to prevent a temporal bias. The top-50 results of the initial retrieval are then re-ranked with a monoT5 cross-encoder model[3] [8] that has been fine-tuned on the MS MARCO passage dataset [45].

To reduce computational complexity, the context length of the model is limited to 512 tokens. Thus, longer web documents need to be split into shorter text passages using a sliding window approach with a length of 400 tokens per passage and a stride of 64 tokens. The passages are scored with monoT5, and finally, the passage-level scores are aggregated after re-ranking. Three aggregation schemes are commonly used [25]:

- The highest score of one of its passages (max passage aggregation),
- the mean score of all of its passages (mean passage aggregation), or
- the mean score of only the top-k ranked passages (k-max average aggregation).

We have submitted one run for each of the three abovementioned aggregation schemes. To find the parameter k for the k-max average aggregation, we ran a Grid Search with $k = 2, 4, \ldots, 20$ on the LongEval data collection from June 2022 that yielded the highest nDCG score [46] at $k = 4$.

3.1.1. Hypotheses

Concerning monoT5 re-ranking, we investigate the following two hypotheses:

Hypothesis 1. *In the setting presented above, the nDCG effectiveness (or nDCG@10, respectively) of monoT5 with max passage score aggregation is significantly higher ($\alpha = 0.05$) than the effectiveness obtained with mean passage aggregation.*

Hypothesis 2. *In the setting presented above, when choosing k such that the nDCG effectiveness (or nDCG@10, respectively) of the k-max average aggregation is maximized, monoT5 with k-max average aggregation yields a significantly higher ($\alpha = 0.05$) nDCG effectiveness (or nDCG@10, respectively) than with max passage or mean passage aggregation.*

Hypothesis 1 builds on the intuition that documents containing relevant passages for a given query are usually considered relevant by users despite possibly also containing irrelevant passages. Hence, the document's relevance would be estimated by the highest relevance of any individual passage from the document. Non-relevant passages should not influence the aggregated scores negatively. However, we question this rather extreme setting and argue that at least a few relevant passages should often be required to make a document relevant. For example, even spam pages could sometimes contain relevant passages by pure chance. Hence, averaging the scores of the best-scoring passages in a document seems intuitive, which we express in Hypothesis 2.

[3] https://huggingface.co/castorini/monot5-base-msmarco

Table 1
Fusion weights after optimizing our weighted sum rank fusion approach towards nDCG@10 on the LongEval January 2023 dataset.

System name	Fusion weight
RankZephyr	0.7
Sparse Cross-Encoder	0.1
ColBERT	0.1
BM25	0.1

3.2. Rank Fusion

Our second approach proposes a weighted rank fusion where we initially retrieve documents with BM25 [23] and re-rank the top-1000 results using a rank fusion model consisting of RankZephyr [33], a sparse cross-encoder [31], ColBERT [32], and BM25. RankZephyr is a model that surpasses GPT-4 [37] performance on several datasets [33] but could also be susceptible to a decline in effectiveness on older data due to its relative novelty. Therefore, other retrieval models are incorporated into the ranking through rank fusion to offset this potential disadvantage and achieve time-resilient effectiveness. We chose the sparse cross-encoder, ColBERT, and BM25 for the rank fusion as they are the most effective models of their respective paradigms (cross-encoder, late interaction, and lexical ranking). Besides this rank fusion, for comparison, we also provide a run that only uses RankZephyr (i.e., no rank fusion).

The rank fusion was implemented as a weighted sum of scores using the Python library ranx.fuse [38]. In ranx.fuse, the scores of all constituent models are computed and optimal weights are assigned to the models' scores based on a given training dataset. Moreover, before the results from different retrieval models can be fused, the document scores are normalized to make them comparable. This step is necessary because the retrieval models use different scales for scoring [47]. We used the standard min-max-normalization, shifting the minimum score to 0 and scaling the maximum score to 1 [47]. A weighted sum was selected as the fusion method, as the weights it assigns to the constituent models' scores are easy to interpret. We optimized the fusion for an optimal nDCG@10 score on the LongEval January 2023 dataset which yielded the weights listed in Table 1.

3.2.1. Hypotheses

Based on our rank fusion approach, we investigate the following hypotheses:

Hypothesis 3. *The differences in the nDCG effectiveness (or nDCG@10, respectively) observed over time are significantly smaller (*$\alpha = 0.05$*) for the rank fusion model described above than for just RankZephyr, the sparse cross-encoder, ColBERT, or BM25 alone.*

Hypothesis 4. *Retrieving documents with the optimized rank fusion model of RankZephyr, the sparse cross-encoder, ColBERT, and BM25, as described above, achieves a significantly higher (*$\alpha = 0.05$*) nDCG effectiveness (or nDCG@10, respectively) than using each of these models alone.*

Hypothesis 3 follows the intuition that a fused model, that combines different retrieval approaches, should be more persistent in its effectiveness over time because some of the systems it combines could compensate for errors that other constituent systems make. Since retrieval systems that do not use temporal-bound training data often achieve a more stable but overall poorer level of effectiveness than neural models, we hypothesize that our rank fusion approach yields consistently higher effectiveness than the single models, as expressed in Hypothesis 4.

3.3. Submitted Runs

To improve the reproducibility of our approaches, the submitted runs are published on TIRA and can be accessed via TIRA.[4] We submitted the following five runs:

Run `galapagos-tortoise-bm25-bo1-pl2-monot5-max` A weighted linear combination of BM25 (with Bo1 query expansion; weight: 2) and PL2 (weight: 1), re-ranked with monoT5.[5] After re-ranking, passages are aggregated by the max passage score aggregation.

Run `galapagos-tortoise-bm25-bo1-pl2-monot5-mean` A weighted linear combination of BM25 (with Bo1 query expansion; weight: 2) and PL2 (weight: 1), re-ranked with monoT5.[5] After re-ranking, passages are aggregated by the mean passage score aggregation.

Run `galapagos-tortoise-bm25-bo1-pl2-monot5-kmax-avg-k-4` A weighted linear combination of BM25 (with Bo1 query expansion; weight: 2) and PL2 (weight: 1), re-ranked with monoT5.[5] After re-ranking, passages are aggregated by the k-max average passage score aggregation with $k = 4$, which yielded the highest nDCG on the LongEval June 2022 dataset.

Run `galapagos-tortoise-wsum` A rank fusion (weighted sum, optimized on the January 2023 dataset) of BM25 (weight: 0.1), the sparse cross-encoder (weight: 0.1), ColBERT (weight: 0.1), and RankZephyr (weight: 0.7) re-ranking after retrieving the top-1000 documents with BM25. The models themselves were not fine-tuned.

Run `galapagos-tortoise-rank-zephyr` Re-ranking BM25's top-1000 documents with RankZephyr.

4. Results

4.1. Neural Re-Ranking with monoT5

Table 2 lists the nDCG and nDCG@10 scores achieved by the three monoT5 variants on the LongEval datasets from January, June, and August 2023. It can be seen that deploying a max passage aggregation yields the highest nDCG and nDCG@10 scores on all three datasets. On the datasets from June and August 2023, the scores achieved by max passage are even significantly higher than the ones obtained with mean passage aggregation. On the January dataset, however, max, 4-max average, and mean passage aggregation behave almost identically. As a result, the p values measured on the January dataset are far away from being significant. The difference in retrieval effectiveness between max passage on the one hand and 4-max average and mean passage on the other hand increase considerably over time.

Furthermore, it seems counter-intuitive that 4-max average passage aggregation performs worse than both max and mean passage aggregation on the January 2023 dataset, given that it is actually a hybrid of the two extremes. It would be interesting to inspect this dataset further to get an intuition on why it behaves fundamentally different than the others.

Re-visiting our hypotheses, we can discard Hypothesis 2 that suspected k-max average aggregation to yield significantly higher nDCG and nDCG@10 scores than the competing passage aggregation schemes. Hypothesis 1, stating that max passage aggregation performs significantly better than mean passage aggregation, deserves a more careful investigation since our experiments convey highly contradictory signals. Taking all three datasets into account, we cannot confirm Hypothesis 1.

[4] Submissions on the January 2023 dataset: https://tira.io/task-overview/ir-lab-padua-2024/longeval-2023-01-20240426-training; submissions on the June 2023 dataset: https://tira.io/task-overview/ir-lab-padua-2024/longeval-2023-06-20240422-training; submissions on the August 2023 dataset: https://tira.io/task-overview/ir-lab-padua-2024/longeval-2023-08-20240422-training

[5] https://huggingface.co/castorini/monot5-base-msmarco

Table 2
Retrieval effectiveness of our monoT5 approach with different score aggregations on the LongEval datasets from January, June, and August 2023. The t test based p values are reported compared to max passage aggregation.

System name	nDCG@10		nDCG	
	value	p value	value	p value
January 2023 dataset				
max passage	**0.209**	—	**0.307**	—
4-max avg. passage	0.208	0.86	0.305	0.41
mean passage	0.209	0.93	0.307	0.82
June 2023 dataset				
max passage	**0.196**	—	**0.260**	—
4-max avg. passage	0.191	0.24	0.257	0.24
mean passage	0.184	0.02	0.253	0.02
August 2023 dataset				
max passage	**0.159**	—	**0.198**	—
4-max avg. passage	0.156	0.07	0.196	0.12
mean passage	0.150	<0.01	0.191	<0.01

Table 3
Retrieval effectiveness decline over time of our monoT5 approach with different score aggregations from the January 2023, to June 2023, to August 2023 datasets.

System name	ΔnDCG@10			ΔnDCG		
	Jan→Jun	Jun→Aug	Jan→Aug	Jan→Jun	Jun→Aug	Jan→Aug
max passage	-0.013	-0.037	-0.049	-0.047	-0.063	-0.110
4-max avg. passage	-0.017	-0.035	-0.052	-0.049	-0.061	-0.110
mean passage	-0.025	-0.034	-0.059	-0.054	-0.062	-0.116

Apart from our specific research questions, we notice a decline in retrieval effectiveness with respect to all three aggregation schemes. Table 3 lists the differences in the nDCG (nDCG@10, respectively) scores that were obtained on the January, June and August 2023 datasets. As can be seen, in each fixed interval the decline in effectiveness is similar for all aggregation schemes. We conclude that on our datasets monoT5 is too sensitive towards temporal changes to make fine-tuning its aggregation a question worth investigating further.

4.2. Rank Fusion

Recall that our rank fusion model that was trained to optimize its nDCG@10 score on the LongEval January 2023 dataset weights RankZephyr with 0.7 and all other models, i.e. the sparse cross-encoder, ColBERT and BM25 with 0.1 each. Table 4 compares the nDCG score (nDCG@10 score, respectively) achieved by our rank fusion approach on the LongEval January, June and August 2023 datasets with the respective scores obtained when using each model purely. On the January and August datasets, the fusion approach outperforms all other systems, although the difference to the scores obtained with RankZephyr is only a slight one. On the June dataset RankZephyr beats our fusion approach by a narrow margin. The ranking of all other systems is stable over all datasets: the sparse cross-encoder scores better than ColBERT and BM25 yields the smallest nDCG and nDCG@10 scores. Since all models we investigated re-rank the top-1000 documents retrieved by BM25, this last observation indicates that neural re-ranking does not deteriorate nDCG scores.

The calculated p values suggest that the difference between our rank fusion approach and RankZephyr is not a significant one. This contradicts the intuition we formulated in Hypothesis 4, but seems plausible,

Table 4
Retrieval effectiveness of our rank fusion approach and its fused runs on the LongEval datasets from January, June, and August 2023. The t test based p values are reported compared to our rank fusion approach.

System name	**nDCG@10**		**nDCG**	
	value	p value	value	p value
January 2023 dataset				
our rank fusion approach	**0.251**	—	**0.355**	—
RankZephyr [33]	0.247	0.07	0.353	0.26
Sparse Cross-Encoder [31]	0.221	<0.01	0.337	<0.01
ColBERT [32]	0.216	<0.01	0.330	<0.01
BM25 [48]	0.199	<0.01	0.317	<0.01
June 2023 dataset				
our rank fusion approach	0.228	—	0.293	—
RankZephyr [33]	**0.228**	0.98	**0.295**	0.34
Sparse Cross-Encoder [31]	0.202	$<$0.01	0.277	$<$0.01
ColBERT [32]	0.183	$<$0.01	0.264	$<$0.01
BM25 [48]	0.166	$<$0.01	0.252	$<$0.01
August 2023 dataset				
our rank fusion approach	**0.180**	—	**0.220**	—
RankZephyr [33]	0.178	0.15	0.219	0.52
Sparse Cross-Encoder [31]	0.169	$<$0.01	0.212	$<$0.01
ColBERT [32]	0.161	$<$0.01	0.206	$<$0.01
BM25 [48]	0.141	$<$0.01	0.191	$<$0.01

Table 5
Retrieval effectiveness decline over time from the January 2023, to June 2023, to August 2023 datasets of our rank fusion approach and its fused runs.

System name	**ΔnDCG@10**			**ΔnDCG**		
	Jan→Jun	Jun→Aug	Jan→Aug	Jan→Jun	Jun→Aug	Jan→Aug
our rank fusion approach	-0.024	-0.048	**-0.072**	-0.062	-0.072	**-0.135**
RankZephyr [33]	-0.019	-0.050	**-0.069**	-0.058	-0.075	**-0.134**
Sparse Cross-Encoder [31]	-0.019	-0.033	**-0.052**	-0.060	-0.064	**-0.125**
ColBERT [32]	-0.033	-0.022	**-0.055**	-0.066	-0.058	**-0.124**
BM25 [48]	-0.033	-0.025	**-0.058**	-0.065	-0.061	**-0.125**

given that in our fusion approach RankZephyr's score gets weighted with 0.7 and hence dominates the model. However, on all datasets the nDCG and nDCG@10 scores of our rank fusion approach are significantly higher than the respective scores of the sparse cross-encoder ColBERT and BM25. Excluding RankZephyr, Hypothesis 4 can therefore be confirmed. However, the fusion model presumably benefits greatly from the effectiveness of RankZephyr.

Table 5 visualizes the differences between nDCG scores (nDCG@10 scores, respectively) on the three collections. Similar to our findings in Subsection 4.1, we witness a temporal decline in effectiveness of all retrieval systems. Moreover, we notice that our highest performing systems, i.e. our rank fusion approach and RankZephyr exhibit the greatest temporal over-all decline as well. We can therefore discard Hypothesis 3 that our rank fusion approach are more stable than the other models.

Inspecting the values in Table 5 further, we compute the pairwise Pearson correlations [49] between the declines of all evaluated systems and visualize the result in Table 6. As can be seen, our systems are split into two camps, within which there is a strong pairwise correlation between the declines: The group of systems with highest effectiveness on the one hand, i.e. our rank fusion approach, RankZephyr

Table 6
Pearson correlation in retrieval effectiveness decline over time of our rank fusion approach and its fused runs.

System name	rank fusion	RankZephyr	Sparse Cross-Encoder	ColBERT	BM25
our rank fusion approach	1	**0.990**	**0.968**	0.693	0.745
RankZephyr [33]	**0.990**	1	**0.925**	0.591	0.653
Sparse Cross-Encoder [31]	**0.968**	**0.925**	1	**0.853**	**0.889**
ColBERT [32]	0.693	0.591	**0.853**	1	**0.996**
BM25 [48]	0.745	0.653	**0.889**	**0.996**	1

and the evaluated sparse cross-encoder. And the group of systems with lower effectiveness on the other, i.e. the sparse cross-encoder, ColBERT and BM25. The sparse cross-encoder provides the link between both camps, as its decline correlates strongly with all systems. This finding is somewhat sobering, because regardless of how different the selected retrieval paradigms are, the decline behaves similarly on all systems, and is most drastic in our most effective systems.

5. Conclusion and Future Work

In this paper, we pursued two different research directions to improve the temporal stability of retrieval systems. First, we experimented with different passage score aggregation schemes for monoT5 re-ranking. We hypothesized that k-max average aggregation with a tuned k should yield a higher nDCG and nDCG@10 effectiveness than max passage aggregation, which in turn should outperform mean passage aggregation. Second, we proposed a weighted rank fusion of RankZephyr, a sparse cross-encoder, ColBERT, and BM25. Here, we expected the rank fusion approach to be both, more effective and more temporally stable than each of the fused models alone.

Regarding neural re-ranking with monoT5 of different aggregation schemes, max passage aggregation indeed outperforms mean passage aggregation with respect to nDCG and nDCG@10, with more significant differences on more recent datasets. Additionally, max passage aggregation was found to be superior to k-max passage, contrary to our hypothesis.

No significant difference was found between the effectiveness of RankZephyr alone and our rank fusion approach. Albeit, the fusion model yielded significantly higher nDCG and nDCG@10 effectiveness compared to BM25, the sparse cross-encoder, and ColBERT. This improvement, however, is likely an effect of the high effectiveness of RankZephyr and its high weight within the fusion model.

We observed that, despite our efforts, the effectiveness of all evaluated retrieval systems declines over time. Moreover, the rates at which nDCG and nDCG@10 scores decrease are highly pairwise correlated between the high-performing rank fusion approach, RankZephyr and the sparse cross-encoder and generally higher than the decline rates of the less effective ColBERT and BM25. Effectiveness and temporal stability seem to work against each other here.

Still, it is contrary to the intuition, that not only the effectiveness of neural re-ranking approaches, but also of lexical models like BM25 declines over time. While the decline in the effectiveness of neural models is usually attributed to the increasingly stale data they were trained on, we lack a good intuition for the temporal decline in BM25's effectiveness. Hence, it would be worthwhile to investigate whether the observed decline in retrieval effectiveness of several basic lexical models is statistically significant over time, to finally distinguish systems with temporal effectiveness decline from those without.

Further research is also needed to explore the effectiveness of rank fusions whose constituent models are equally well-performing and more diverse in the conceptual retrieval approach they implement. Conducting a larger study with diverse fusion candidates could hopefully lead to the development of effective and temporally stable hybrid models.

Our research contributes to the understanding of long-term stability in retrieval systems, providing insights into the performance of various passage score aggregation schemes with monoT5 and rank fusion methods. Despite observing a general decline in effectiveness over time, our findings highlight the potential of hybrid models that integrate both neural and lexical approaches and show that further researches in optimized aggregation techniques or fusion strategies with more diverse candidates can lead to enhanced long-term retrieval performance.

References

[1] I. Matveeva, C. Burges, T. Burkard, A. Laucius, L. Wong, High accuracy retrieval with multiple nested ranker, in: E. N. Efthimiadis, S. T. Dumais, D. Hawking, K. Järvelin (Eds.), Proceedings of SIGIR 2006, ACM, 2006, pp. 437–444. doi:10.1145/1148170.1148246.

[2] L. Wang, J. Lin, D. Metzler, A cascade ranking model for efficient ranked retrieval, in: W. Ma, J. Nie, R. Baeza-Yates, T. Chua, W. B. Croft (Eds.), Proceedings of SIGIR 2011, ACM, 2011, pp. 105–114. doi:10.1145/2009916.2009934.

[3] N. Asadi, J. Lin, Effectiveness/efficiency tradeoffs for candidate generation in multi-stage retrieval architectures, in: G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, T. Sakai (Eds.), The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013, ACM, 2013, pp. 997–1000. doi:10.1145/2484028.2484132.

[4] R. Chen, L. Gallagher, R. Blanco, J. S. Culpepper, Efficient cost-aware cascade ranking in multi-stage retrieval, in: N. Kando, T. Sakai, H. Joho, H. Li, A. P. de Vries, R. W. White (Eds.), Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017, ACM, 2017, pp. 445–454. doi:10.1145/3077136.3080819.

[5] J. M. Mackenzie, J. S. Culpepper, R. Blanco, M. Crane, C. L. A. Clarke, J. Lin, Query driven algorithm selection in early stage retrieval, in: Y. Chang, C. Zhai, Y. Liu, Y. Maarek (Eds.), Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018, ACM, 2018, pp. 396–404. doi:10.1145/3159652.3159676.

[6] R. F. Nogueira, W. Yang, K. Cho, J. Lin, Multi-stage document ranking with BERT (2019). URL: http://arxiv.org/abs/1910.14424. arXiv:1910.14424.

[7] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, L. Deng, MS MARCO: A human generated machine reading comprehension dataset, in: T. R. Besold, A. Bordes, A. S. d'Avila Garcez, G. Wayne (Eds.), Proceedings of CoCo@NIPS 2016, volume 1773 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2016. URL: https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

[8] R. Pradeep, R. F. Nogueira, J. Lin, The Expando-Mono-Duo design pattern for text ranking with pretrained sequence-to-sequence models (2021). URL: https://arxiv.org/abs/2101.05667. arXiv:2101.05667.

[9] T. Linjordet, K. Balog, Sanitizing synthetic training data generation for question answering over knowledge graphs, in: K. Balog, V. Setty, C. Lioma, Y. Liu, M. Zhang, K. Berberich (Eds.), ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020, ACM, 2020, pp. 121–128. doi:10.1145/3409256.3409836.

[10] K. Krishna, A. Roy, M. Iyyer, Hurdles to progress in long-form question answering, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 4940–4957. doi:10.18653/v1/2021.naacl-main.393.

[11] M. Fröbe, C. Akiki, M. Potthast, M. Hagen, How train-test leakage affects zero-shot retrieval, in: D. Arroyuelo, B. Poblete (Eds.), String Processing and Information Retrieval - 29th International Symposium, SPIRE 2022, Concepción, Chile, November 8-10, 2022, Proceedings, volume 13617 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 147–161. doi:10.1007/978-3-031-20643-6_11.

[12] E. G. Altmann, J. B. Pierrehumbert, A. E. Motter, Beyond word frequency: Bursts, lulls, and scaling

in the temporal distributions of words, CoRR abs/0901.2349 (2009). URL: http://arxiv.org/abs/0901.2349. arXiv:0901.2349.

[13] W. Labov, Principles of linguistic change, volume volume 3, John Wiley and Sons, 2011.

[14] S. Longpre, G. Yauney, E. Reif, K. Lee, A. Roberts, B. Zoph, D. Zhou, J. Wei, K. Robinson, D. Mimno, D. Ippolito, A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, CoRR abs/2305.13169 (2023). URL: https://doi.org/10.48550/arXiv.2305.13169. doi:10.48550/ARXIV.2305.13169. arXiv:2305.13169.

[15] R. G. Reddy, B. Iyer, M. A. Sultan, R. Zhang, A. Sil, V. Castelli, R. Florian, S. Roukos, Synthetic target domain supervision for open retrieval QA, CoRR abs/2204.09248 (2022). doi:10.48550/arXiv.2204.09248. arXiv:2204.09248.

[16] R. Alkhalifa, E. Kochkina, A. Zubiaga, Building for tomorrow: Assessing the temporal persistence of text classifiers, Inf. Process. Manag. 60 (2023) 103200. doi:10.1016/J.IPM.2022.103200.

[17] R. Ren, Y. Qu, J. Liu, X. Zhao, Q. Wu, Y. Ding, H. Wu, H. Wang, J. Wen, A thorough examination on zero-shot dense retrieval, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 15783–15796. doi:10.18653/V1/2023.FINDINGS-EMNLP.1057.

[18] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, T. Fink, P. Galuščáková, G. Gonzalez-Saez, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, A. Zubiaga, Overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2024.

[19] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. Espinosa-Anke, T. Fink, P. Galuščáková, G. Gonzalez-Saez, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, A. Zubiaga, Extended overview of the CLEF 2024 LongEval Lab on Longitudinal Evaluation of Model Performance, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.

[20] R. F. Nogueira, W. Yang, J. Lin, K. Cho, Document expansion by query prediction, CoRR abs/1904.08375 (2019). URL: http://arxiv.org/abs/1904.08375. arXiv:1904.08375.

[21] G. Amati, Probability models for information retrieval based on divergence from randomness, Ph.D. thesis, University of Glasgow, UK, 2003. URL: http://theses.gla.ac.uk/1570/.

[22] G. Amati, C. J. van Rijsbergen, Probabilistic models of information retrieval based on measuring the divergence from randomness, ACM Trans. Inf. Syst. 20 (2002) 357–389. doi:10.1145/582415.582416.

[23] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[24] C. G. Figuerola, J. L. A. Berrocal, Á. F. Z. Rodríguez, Segmentation of web documents and retrieval of useful passages, in: C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras, D. Santos (Eds.), Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, volume 5152 of *Lecture Notes in Computer Science*, Springer, 2007, pp. 732–736. URL: https://doi.org/10.1007/978-3-540-85760-0_93. doi:10.1007/978-3-540-85760-0_93.

[25] Z. Dai, J. Callan, Deeper text understanding for IR with contextual neural language modeling, in: B. Piwowarski, M. Chevalier, É. Gaussier, Y. Maarek, J. Nie, F. Scholer (Eds.), Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, ACM, 2019, pp. 985–988. URL: https://doi.org/10.1145/

3331184.3331303. doi:10.1145/3331184.3331303.

[26] E. A. Fox, J. A. Shaw, Combination of multiple searches, in: D. K. Harman (Ed.), Proceedings of TREC 1993, volume 500-215 of *NIST Special Publication*, NIST, 1993, pp. 243–252. URL: http://trec.nist.gov/pubs/trec2/papers/ps/vpi.ps.

[27] J. Lee, Analyses of multiple evidence combination, in: N. J. Belkin, A. D. Narasimhalu, P. Willett, W. R. Hersh, F. Can, E. M. Voorhees (Eds.), Proceedings of SIGIR 1997, ACM, 1997, pp. 267–276. doi:10.1145/258525.258587.

[28] J. A. Aslam, M. H. Montague, Models for metasearch, in: W. B. Croft, D. J. Harper, D. H. Kraft, J. Zobel (Eds.), Proceedings of SIGIR 2001, ACM, 2001, pp. 275–284. doi:10.1145/383952.384007.

[29] D. Lillis, F. Toolan, R. W. Collier, J. Dunnion, ProbFuse: a probabilistic approach to data fusion, in: E. N. Efthimiadis, S. T. Dumais, D. Hawking, K. Järvelin (Eds.), Proceedings of SIGIR 2006, ACM, 2006, pp. 139–146. doi:10.1145/1148170.1148197.

[30] G. V. Cormack, C. L. A. Clarke, S. Büttcher, Reciprocal rank fusion outperforms condorcet and individual rank learning methods, in: J. Allan, J. A. Aslam, M. Sanderson, C. Zhai, J. Zobel (Eds.), Proceedings of SIGIR 2009, ACM, 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[31] F. Schlatt, M. Fröbe, M. Hagen, Investigating the Effects of Sparse Attention on Cross-Encoders, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Proceedings of ECIR 2024, volume 14608 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 173–190. doi:10.1007/978-3-031-56027-9_11.

[32] O. Khattab, M. Zaharia, ColBERT: Efficient and effective passage search via contextualized late interaction over BERT (2020). URL: https://arxiv.org/abs/2004.12832. arXiv:2004.12832.

[33] R. Pradeep, S. Sharifymoghaddam, J. Lin, RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! (2023). doi:10.48550/arXiv.2312.02724. arXiv:2312.02724.

[34] R. F. Nogueira, K. Cho, Passage re-ranking with BERT, CoRR abs/1901.04085 (2019). URL: http://arxiv.org/abs/1901.04085. arXiv:1901.04085.

[35] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, O. Frieder, Efficient document re-ranking for transformers by precomputing term representations, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 49–58. doi:10.1145/3397271.3401093.

[36] H. Scells, S. Zhuang, G. Zuccon, Reduce, reuse, recycle: Green information retrieval research, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 2825–2837. doi:10.1145/3477495.3531766.

[37] OpenAI, Gpt-4 technical report, 2024. arXiv:2303.08774.

[38] E. Bassani, L. Romelli, ranx.fuse: A Python library for metasearch, in: M. A. Hasan, L. Xiong (Eds.), Proceedings of CIKM 2022, ACM, 2022, pp. 4808–4812. doi:10.1145/3511808.3557207.

[39] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval, in: Proceedings of CIKM 2021, ACM, 2021, pp. 4526–4533. doi:10.1145/3459637.3482013.

[40] C. Macdonald, N. Tonellotto, Declarative experimentation in information retrieval using PyTerrier, in: Proceedings of ICTIR 2020, 2020.

[41] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), Proceedings of SIGIR 2021, ACM, 2021, pp. 2429–2436. doi:10.1145/3404835.3463254.

[42] M. Fröbe, J. H. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The information retrieval experiment platform, in: Proceedings of SIGIR 2023, ACM, 2023, pp. 2826–2836. doi:10.1145/3539618.3591888.

[43] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: Proceedings of ECIR 2023, Lecture Notes in Computer Science, Springer, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[44] P. Galuscáková, R. Deveaud, G. G. Sáez, P. Mulhem, L. Goeuriot, F. Piroi, M. Popel, LongEval-retrieval: French-english dynamic test collection for continuous web search evaluation, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of SIGIR 2023, ACM, 2023, pp. 3086–3094. doi:10.1145/3539618.3591921.

[45] R. F. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of EMNLP 2020, volume EMNLP 2020 of *Findings of ACL*, ACL, 2020, pp. 708–718. doi:10.18653/v1/2020.findings-emnlp.63.

[46] K. Järvelin, J. Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Trans. Inf. Syst. 20 (2002) 422–446. URL: http://doi.acm.org/10.1145/582415.582418. doi:10.1145/582415.582418.

[47] M. H. Montague, J. A. Aslam, Relevance score normalization for metasearch, in: Proceedings of the 2001 ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA, November 5-10, 2001, ACM, 2001, pp. 427–433. doi:10.1145/502585.502657.

[48] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: W. B. Croft, C. J. van Rijsbergen (Eds.), Proceedings of SIGIR 1994, ACM/Springer, 1994, pp. 232–241. doi:10.1007/978-1-4471-2099-5_24.

[49] K. Pearson, Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.