Overview of the (Ongoing) ReNeuIR 2024 Shared Task

SIGIR 2024, July 14–18, Washington D.C., USA

Maik Fröbe, Joel Mackenzie, Bhaskar Mitra, Franco Maria Nardini, and Martin Potthast

University of Jena University of Queensland Microsoft Research

ISTI-CNR University of Kassel

@ReNeuIRWorkshop reneuir.org

Is Our Picture of IR Evaluation Complete?

IR has a strong focus on effectiveness

- 1966: Cranfield test collection

We pay attention to the details:

- Significance tests?
- Multiple tests?
- MRR: Pro or con?

Efficiency often missing in the picture

Is Our Picture of IR Evaluation Complete?

IR has a strong focus on effectiveness

- 1966: Cranfield test collection

We pay attention to the details:

- Significance tests?
- Multiple tests?
- MRR: Pro or con?

Efficiency often missing in the picture

Goal of ReNeuIR: Complete the Evaluation Picture

Enable evaluation measures that combine efficiency and effectiveness

Enabling An Holistic Picture of Efficiency and Effectiveness

How Could the Complete Picture Look Like?

- We “outsource” effectiveness to others
- We want to enable as much interpretations of efficiency as possible
- Very rough efficiency classes first, i.e., using log scale for efficiency first

Painting the Picture(s) takes Multiple Years

Where do we come from

- ReNeuIR 2022: What to measure? (power, emissions, elapsed time, . . .)
- ReNeuIR 2023: Develop methodology for a shared task

Painting the Picture(s) takes Multiple Years

Where do we come from

- ReNeuIR 2022: What to measure? (power, emissions, elapsed time, . . .)
- ReNeuIR 2023: Develop methodology for a shared task

Where we currently are

- Running the shared task: put as much dots on the figure as possible
- Produce a parallel corpus: Run file + telemetry

Painting the Picture(s) takes Multiple Years

Where do we come from

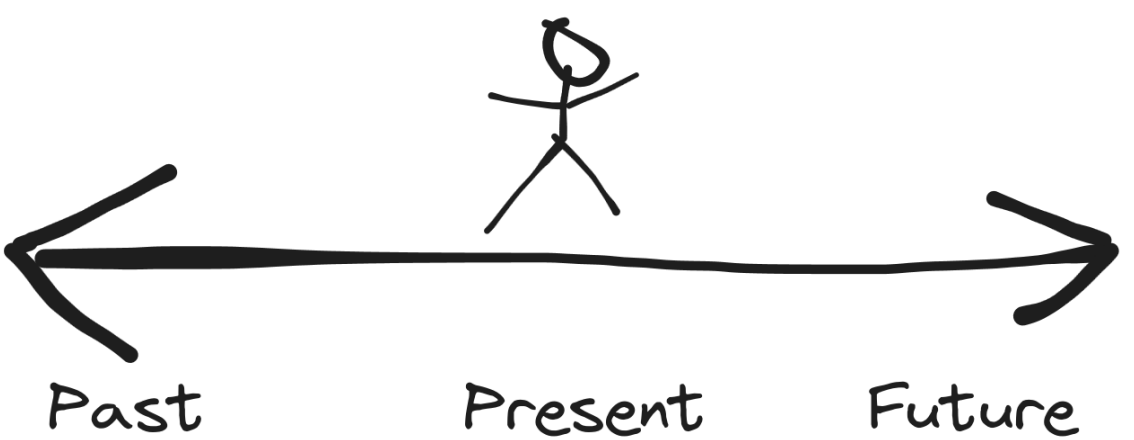
- ReNeuIR 2022: What to measure? (power, emissions, elapsed time, . . .)
- ReNeuIR 2023: Develop methodology for a shared task

Where we currently are

- Running the shared task: put as much dots on the figure as possible
- Produce a parallel corpus: Run file + telemetry

Where we might go to

- ReNeuIR 2025: Collect, compare, and discuss new proposed measures
- Ideally while using/enriching the parallel corpus



Overview of the ReNeuIR 2024 Shared Task: Corpora

MS Marco Passage v1 as Collection

- Many trained systems exist already
- We vary the loads of documents and queries

Overview of the ReNeuIR 2024 Shared Task: Corpora

MS Marco Passage v1 as Collection

- Many trained systems exist already
- We vary the loads of documents and queries

Varying Document Load

- Starting point: 97 queries from TREC DL 19/20 pool of all submitted runs
- Pooling for different dataset sizes

Pool	Queries	Documents
10	97	6965
100	97	68261
1000	97	543311

Overview of the ReNeuIR 2024 Shared Task: Corpora

MS Marco Passage v1 as Collection

- Many trained systems exist already
- We vary the loads of documents and queries

Varying Document Load

- Starting point: 97 queries from TREC DL 19/20 pool of all submitted runs
- Pooling for different dataset sizes

Pool	Queries	Documents
10	97	6965
100	97	68261
1000	97	543311

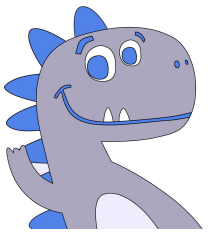
Varying Query Load

- Starting point: MS MARCO Dev/small dataset
- Top-500 pool as set of documents
- Varying number of queries

Queries	Documents
100	2 314 745
1000	2 314 745
6980	2 314 745

Overview of the ReNeuIR 2024 Shared Task: Submission System

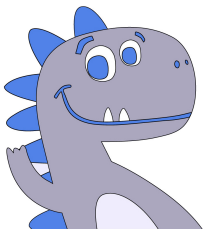
TIRA/TIREx for Software Submissions



- Approaches implemented against ir_datasets
  - We can inject different data loads without data wrangling
- Upload docker image
  - Rule of thumb: Putting the code + models somewhere is enough
  - We have prepared Github actions, workflow (runtime around 10 minutes):
    - Build Docker image
    - Test Docker image on tiny dataset
    - Upload Docker image to TIRA
- Executed within TIRA sandbox: No internet for improved reproducibility

Overview of the ReNeuIR 2024 Shared Task: Submission System

TIRA/TIREx for Software Submissions



- Approaches implemented against ir_datasets
  - We can inject different data loads without data wrangling
- Upload docker image
  - Rule of thumb: Putting the code + models somewhere is enough
  - We have prepared Github actions, workflow (runtime around 10 minutes):
    - Build Docker image
    - Test Docker image on tiny dataset
    - Upload Docker image to TIRA
- Executed within TIRA sandbox: No internet for improved reproducibility

Hardware

- TIRA uses kubernetes with 144 nodes (including 24 A100, 24 GTX 1080)
- All telemetry measurements pinned to same host with same specs
  - 1 A100 GPU with 40GB
  - 5 CPU cores + 50GB RAM
  - Timeout: 24 hours

Overview of the ReNeuIR 2024 Shared Task: Telemetry

Overview of Telemetry

- We monitor CPU/GPU utilization, Memory usage, etc. during execution
- Parallel corpus of telemetry + run files archived on Zenodo
- Simplified access via TIRA Python API

```python
nvidia_smi_log = tira.profiling.raw_telemetry(
    'reneuir-2024/tinyfsu/tiny-fsu-bert',
    dataset='dl-top-1000-docs-20240701-training',
    resource='nvidia-smi.log'
)

print(nvidia_smi_log[:5000])
```

```
==============NVSMI LOG==============

Timestamp                                 : Sun Jul  7 11:31:34 2024
Driver Version                            : 545.29.06
CUDA Version                              : 12.3

Attached GPUs                             : 1
GPU 00000000:41:00.0
    FB Memory Usage
        Total                             : 40960 MiB
        Reserved                          : 621 MiB
        Used                              : 4 MiB
        Free                              : 40333 MiB
    BAR1 Memory Usage
        Total                             : 65536 MiB
        Used                              : 1 MiB
        Free                              : 65535 MiB
    Conf Compute Protected Memory Usage
        Total                             : 0 MiB
```

Overview of the ReNeuIR 2024 Shared Task: Telemetry

Overview of Telemetry

- We monitor CPU/GPU utilization, Memory usage, etc. during execution
- Parallel corpus of telemetry + run files archived on Zenodo
- Simplified access via TIRA Python API

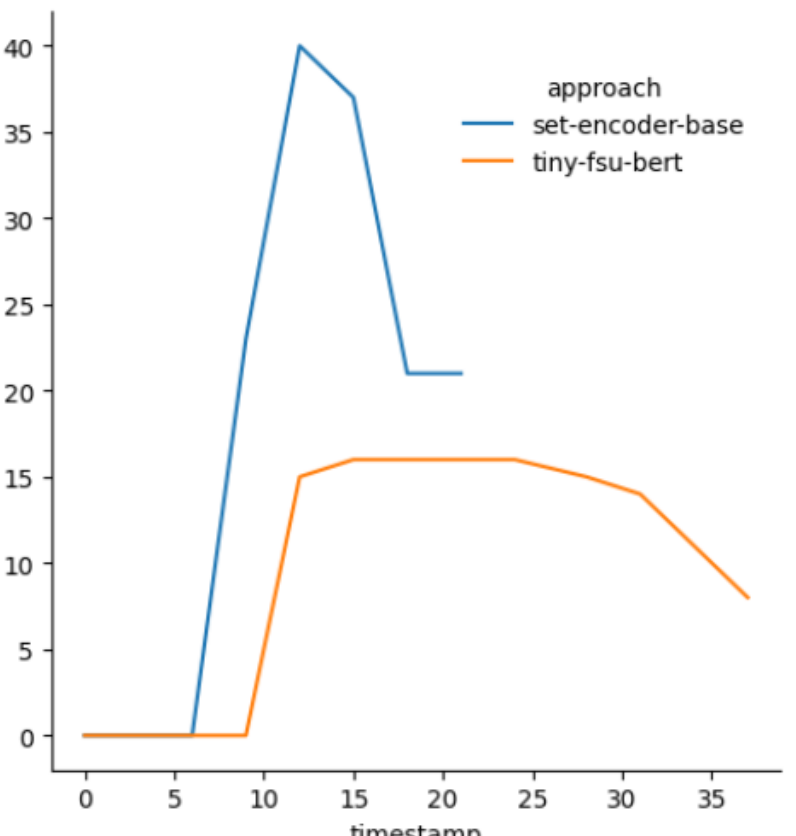
```
set_encoder_profiling = tira.profiling.from_submission(
    'reneuir-2024/fschlatt/set-encoder-base',
    dataset='dl-top-1000-docs-20240701-training',
    return_pd=True
)

set_encoder_profiling.sort_values('timestamp', ascending=True).head(15)
```

	timestamp	key	value
0	0.0	ps_cpu	0.4
1	0.0	ps_vsz	2975140.0
2	0.0	ps_rss	239616.0
25	0.0	gpu_utilization	0 %
24	0.0	gpu_memory_used	4 MiB
3	3.0	ps_cpu	116.3
4	3.0	ps_vsz	3306488.0
5	3.0	ps_rss	513520.0
27	3.0	gpu_utilization	0 %
26	3.0	gpu_memory_used	4 MiB

Overview of the ReNeuIR 2024 Shared Task: Telemetry

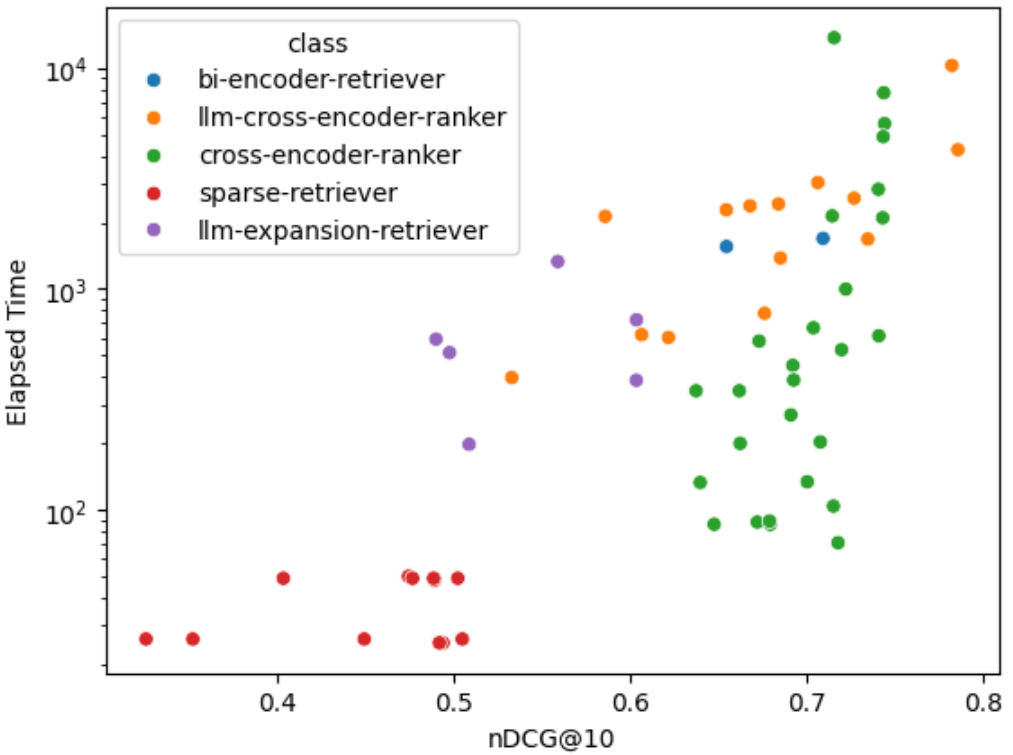
Overview of Telemetry

- We monitor CPU/GPU utilization, Memory usage, etc. during execution
- Parallel corpus of telemetry + run files archived on Zenodo
- Simplified access via TIRA Python API



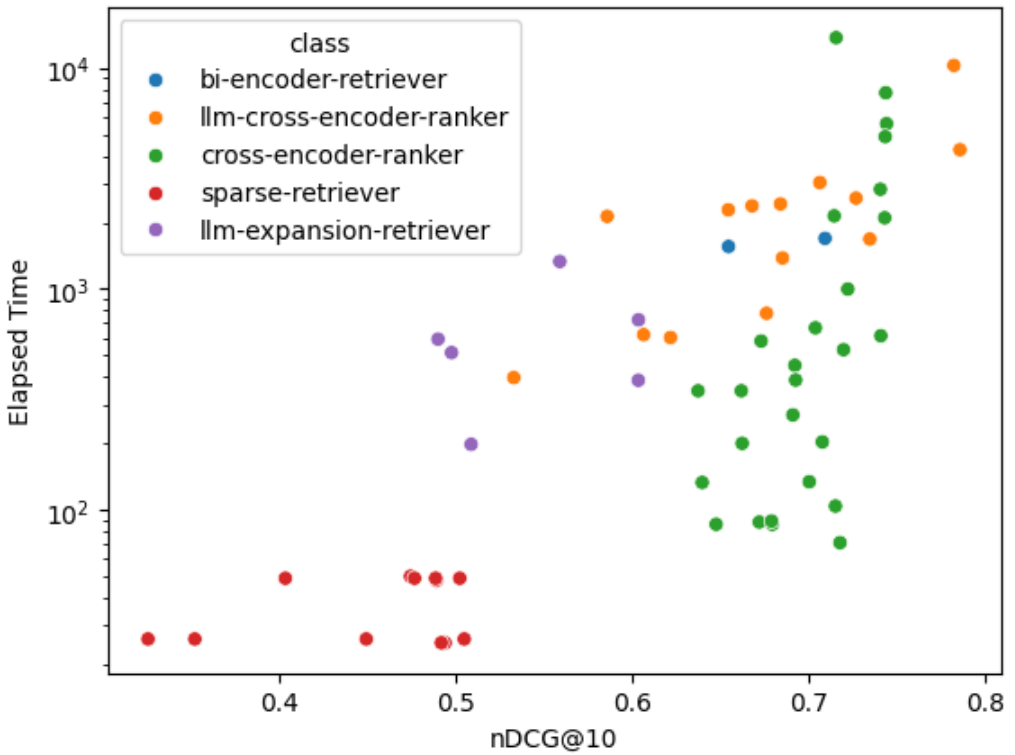
Overview of the ReNeuIR 2024 Shared Task: Systems

Overview of Systems so far



Overview of the ReNeuIR 2024 Shared Task: Systems

Overview of Systems so far

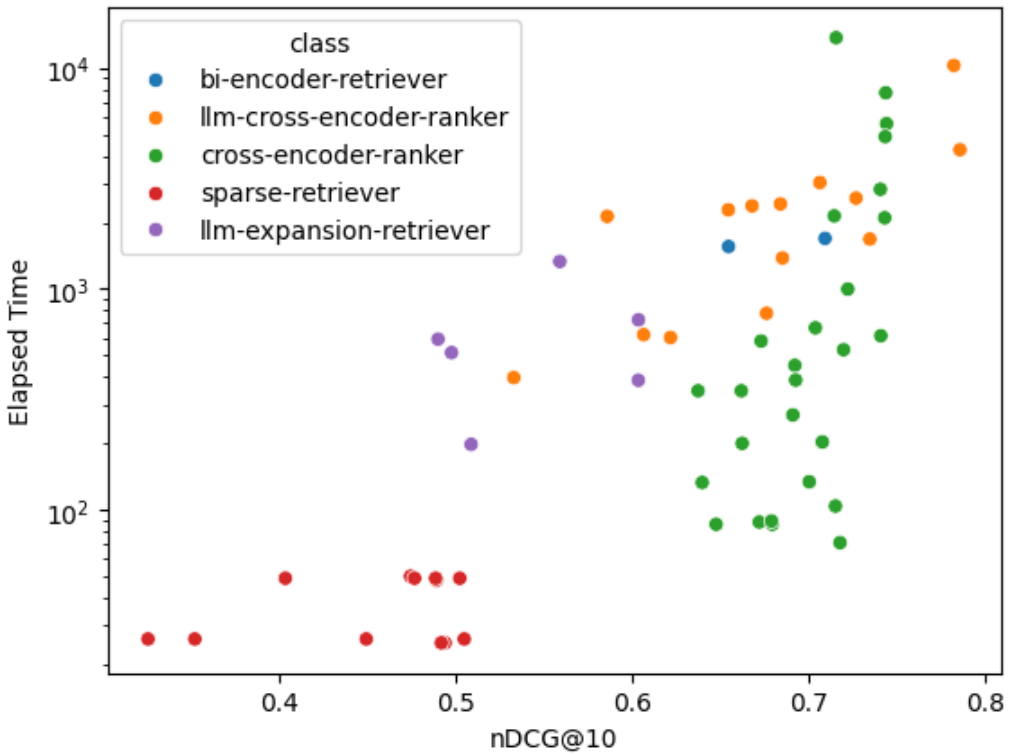


61 Systems (DL 19/20)

System	El. Time	nDCG@10
BM25 (PyTerrier)	48	0.49
BM25 (Anserini)	25	0.49

Overview of the ReNeuIR 2024 Shared Task: Systems

Overview of Systems so far

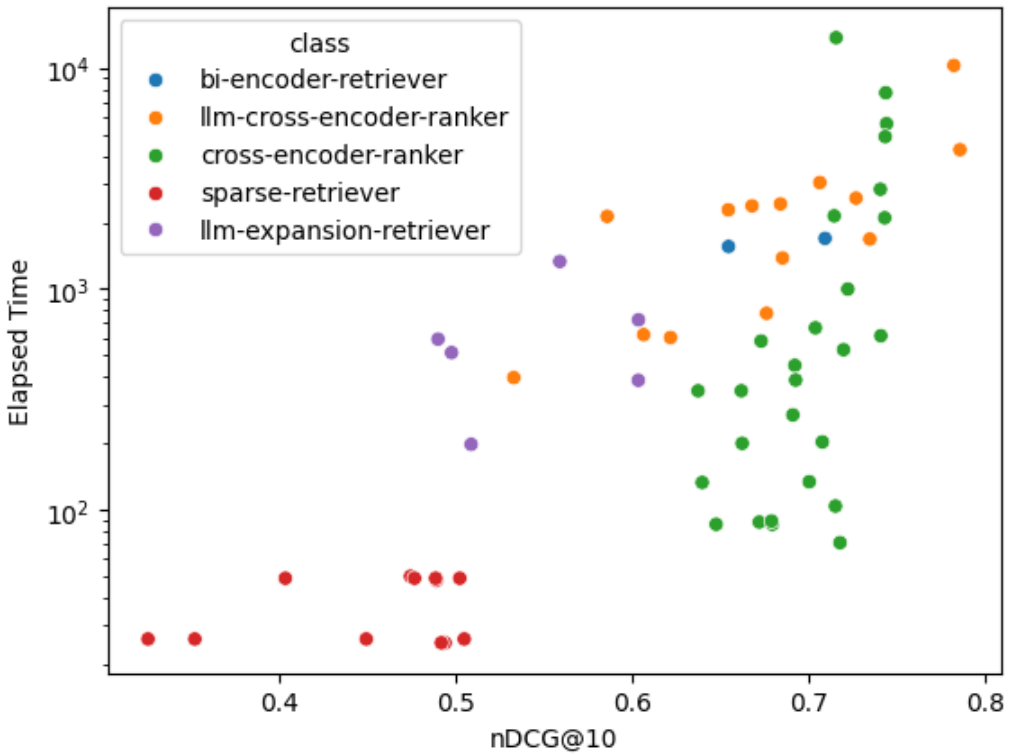


61 Systems (DL 19/20)

System	El. Time	nDCG@10
BM25 >> TinyBERT	86	0.68
BM25 >> SetEncoder	71	0.72
BM25 (PyTerrier)	48	0.49
BM25 (Anserini)	25	0.49

Overview of the ReNeuIR 2024 Shared Task: Systems

Overview of Systems so far

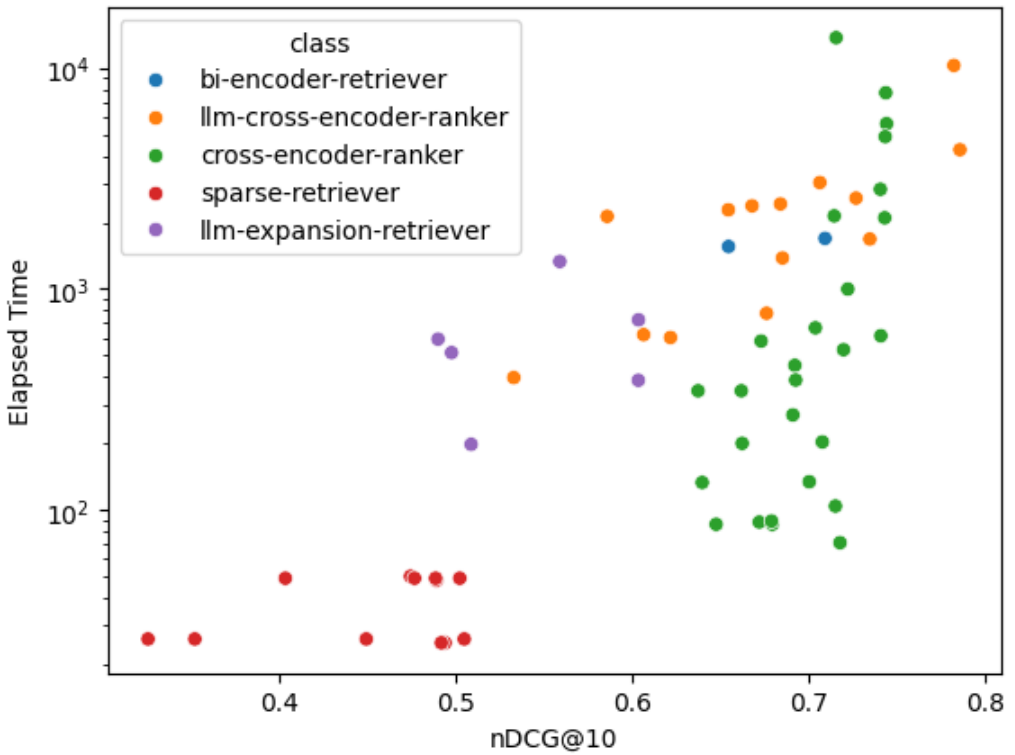


61 Systems (DL 19/20)

System	El. Time	nDCG@10
PLAID-X >> RankZephyr	4272	0.79
BM25 >> RankLlama	1683	0.73
Plaid-X	1695	0.71
BM25 >> TinyBERT	86	0.68
BM25 >> SetEncoder	71	0.72
BM25 (PyTerrier)	48	0.49
BM25 (Anserini)	25	0.49

Overview of the ReNeuIR 2024 Shared Task: Systems

Overview of Systems so far



61 Systems (DL 19/20)

System	El. Time	nDCG@10
Fusion-T5 >> RankZephyr	10285	0.78
PLAID-X >> RankZephyr	4272	0.79
BM25 >> RankLlama	1683	0.73
Plaid-X	1695	0.71
BM25 >> TinyBERT	86	0.68
BM25 >> SetEncoder	71	0.72
BM25 (PyTerrier)	48	0.49
BM25 (Anserini)	25	0.49

Submissions still open :)

Overview of the ReNeuIR 2024 Shared Task

Conclusions and Future Work

Goal of the ReNeuIR 2024 shared task

- Foster development of new performance measures that incorporate efficiency and effectiveness simultaneously
- Collect many diverse systems to cover diverse efficiency / effectiveness tradeoffs

Overview of the ReNeuIR 2024 Shared Task

Conclusions and Future Work

Goal of the ReNeuIR 2024 shared task

- Foster development of new performance measures that incorporate efficiency and effectiveness simultaneously
- Collect many diverse systems to cover diverse efficiency / effectiveness tradeoffs

Results

- Parallel corpus: Runs + Telemetry on Zenodo
- 61 systems at the moment
- 6 MS MARCO subcorpora varying query/document loads

Overview of the ReNeuIR 2024 Shared Task

Conclusions and Future Work

Goal of the ReNeuIR 2024 shared task

- Foster development of new performance measures that incorporate efficiency and effectiveness simultaneously
- Collect many diverse systems to cover diverse efficiency / effectiveness tradeoffs

Results

- Parallel corpus: Runs + Telemetry on Zenodo
- 61 systems at the moment
- 6 MS MARCO subcorpora varying query/document loads

Future iterations

- Some submissions still being finalized. Call still open, please submit :)
- Develop, test, compare holistic measures
- Support to research how to measure efficiency and trade it off for effectiveness

Overview of the ReNeuIR 2024 Shared Task

Conclusions and Future Work

Goal of the ReNeuIR 2024 shared task

- Foster development of new performance measures that incorporate efficiency and effectiveness simultaneously
- Collect many diverse systems to cover diverse efficiency / effectiveness tradeoffs

Thank you!

Results

- Parallel corpus: Runs + Telemetry on Zenodo
- 61 systems at the moment
- 6 MS MARCO subcorpora varying query/document loads

Future iterations

- Some submissions still being finalized. Call still open, please submit :)
- Develop, test, compare holistic measures
- Support to research how to measure efficiency and trade it off for effectiveness