Evaluating Temporal Persistence Using Replicability Measures

LongEval at CLEF | Thessaloniki | 18-21 September 2023

***Jüri Keller**, Timo Breuer and Philipp Schaer*

Technology Arts Sciences TH Köln

Outline

- LongEval Sub-collections
- Retrieval Systems
- Longitudinal Evaluation as Replicability
- Conclusion and Outlook

Evaluating Temporal Persistence Using Replicability Measures

Notebook for the LongEval Lab at CLEF 2023

Jüri Keller[1,*], Timo Breuer[1] and Philipp Schaer[1]

[1] TH Köln (University of Applied Sciences), Claudiusstr. 1, Cologne, 50678, Germany

Abstract
In real-world Information Retrieval (IR) experiments, the Evaluation Environment (EE) is exposed to constant change. Documents are added, removed, or updated, and the information need and the search behavior of users is evolving. Simultaneously, IR systems are expected to retain a consistent quality. The LongEval Lab seeks to investigate the longitudinal persistence of IR systems, and in this work, we describe our participation. We submitted runs of five advanced retrieval systems, namely a Reciprocal Rank Fusion (RRF) approach, ColBERT, monoT5, Doc2Query, and E5, to both sub-tasks. Further, we cast the longitudinal evaluation as a replicability study to better understand the temporal change observed. As a result, we quantify the persistence of the submitted runs and see great potential in this evaluation method.

Keywords
web search, longitudinal evaluation, continuous evaluation, replicability

1. Introduction

This paper describes our contribution to the CLEF 2023 LongEval Lab [1].[1] The lab seeks to investigate the temporal persistence of retrieval systems. It, therefore, provides a first-of-its-kind web retrieval collection with three sub-collections from different points in time [2]. We participated in the retrieval task by providing runs of five systems to both sub-task.

A retrieval system's Evaluation Environment (EE) is under constant change. Not only but especially web retrieval systems are exposed to this due to the dynamic nature of the web. Documents, i.e., websites, get created, updated, or created [3, 4]. But besides the evolving collection, all other aspects of an EE underlay change as well, from the information need and search behavior of the users [5] all the way to the evolving language itself [6]. These changes raise questions about the persistence and generalizability of IR system effectiveness evaluations.

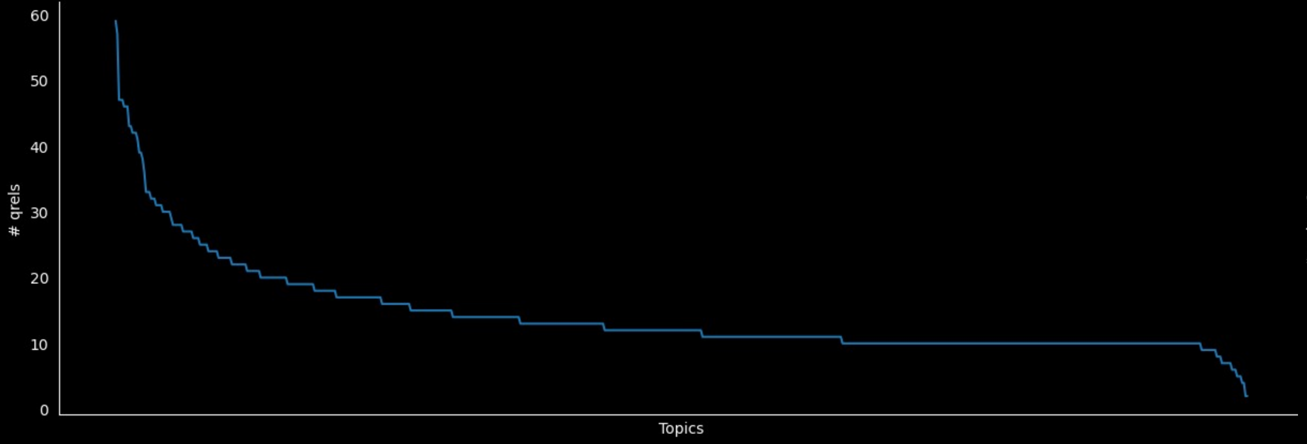
By requiring a temporarily reliable system to perform consistently over time, evaluating this can be understood as a replicability task. Oriented at the ACM definition of replicability[2], the

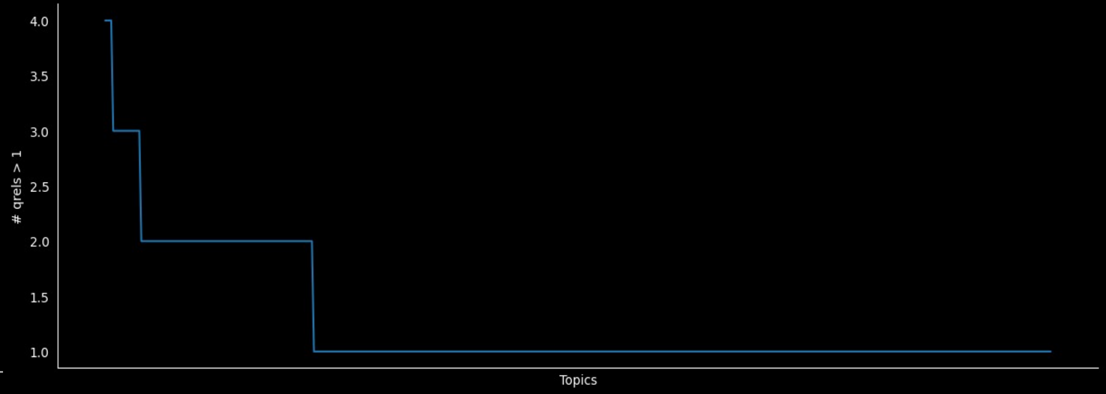
CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
*Corresponding author.
jueri.keller@smail.th-koeln.de (J. Keller); timo.breuer@th-koeln.de (T. Breuer); philipp.schaer@th-koeln.de (P. Schaer)
0000-0002-9392-8646 (J. Keller); 0000-0002-1765-2449 (T. Breuer); 0000-0002-8817-4632 (P. Schaer)

CEUR Workshop Proceedings (CEUR-WS.org)
[1] https://clef-longeval.github.io
[2] https://www.acm.org/publications/policies/artifact-review-and-badging-current

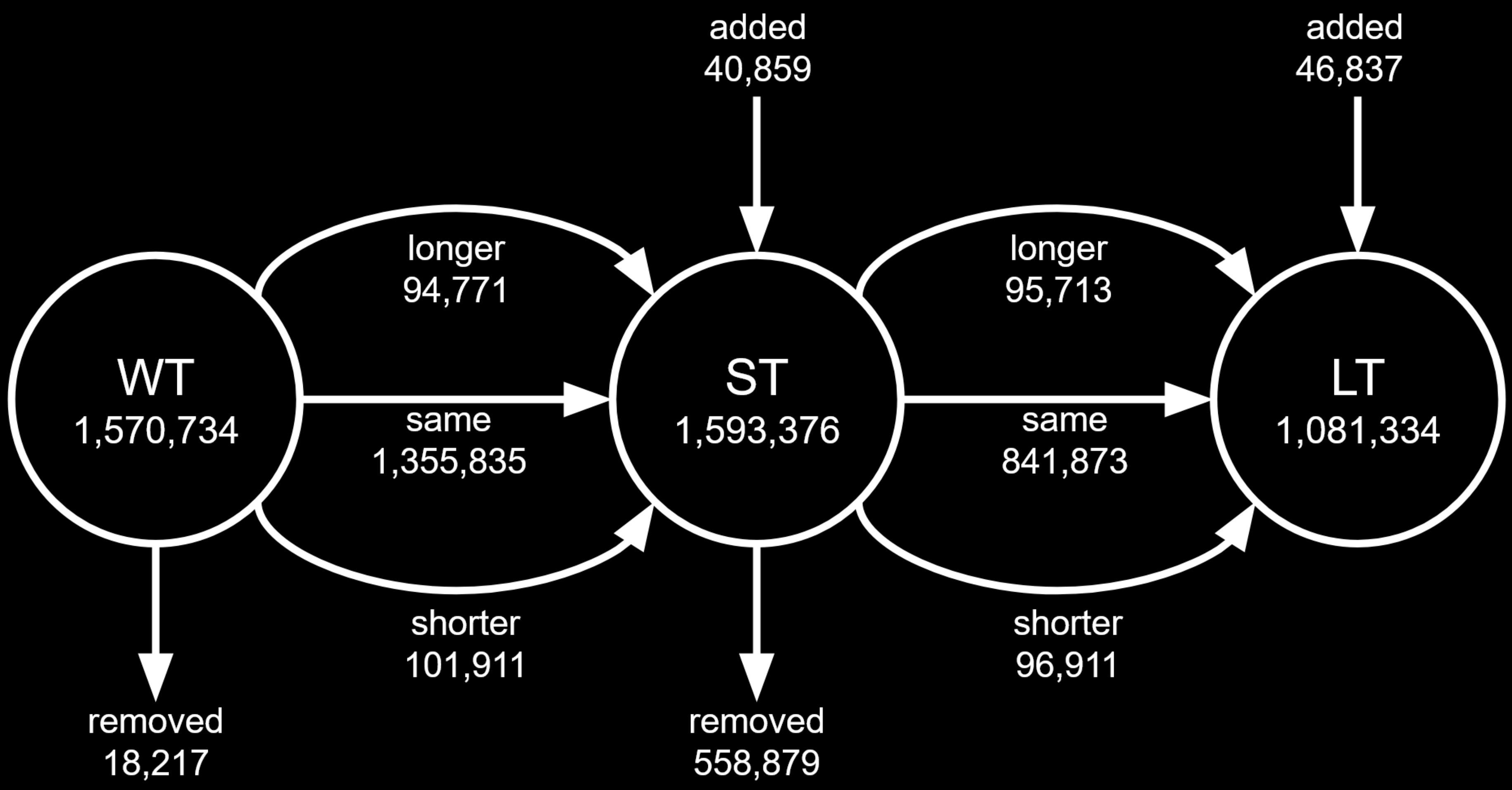
LongEval Sub-collections

- French and automated English translations
- >750 queries from various topics
  - 124 *core queries*
- Over 1.5 million websites
  - 1,011,613 *core websites*
- Qrels from Cascade Click Mode





LongEval Sub-collections



Retrieval Systems

- Submitted to both sub-tasks, ST and LT
- 5 state-of-the-art systems
  - Not adapted to the dataset
  - Not adapted for Temporal IR
  - Allow to set results in a broader context
- Synthesized from various systems

- RRF
- ColBERT
- MonoT5
- Doc2Query
- $E5_{base}$

Retrieval Systems

RRF

- Reciprocal Rank Fusion (RRF) of *BM25+Bo1, DFR,* and *PL2*
- Implemented through *PyTerrier and Ranx*
- fast and computationally inexpensive

Retrieval Systems

ColBERT

- Late Interaction
- Top 1k BM25 results re-ranked
- Zero-shot, trained on MS Marco
- Implemented through *PyTerrier*

Retrieval Systems

MonoT5

- Generative language model for ranking
- Top 1k BM25 results re-ranked
- Based on the first 512 sub-word tokens of a website
- Zero-shot, trained on MS Marco
- Implemented through *PyTerrier*

Retrieval Systems

Doc2Query

- Generative language model for indexing
- Expand each website with ten potential queries
- Based on the first 512 sub-word tokens of a website
- Zero-shot, trained on MS Marco
- Implemented through *PyTerrier*

Retrieval Systems

$E5_{base}$

- Dense retrieval
- Based on the first 512 sub-word tokens of a website
- Zero-shot, trained on CCPairs
- Implemented through *Hugging Face* and *Faiss*

Longitudinal Evaluations

Few statistical significant improvements in WT

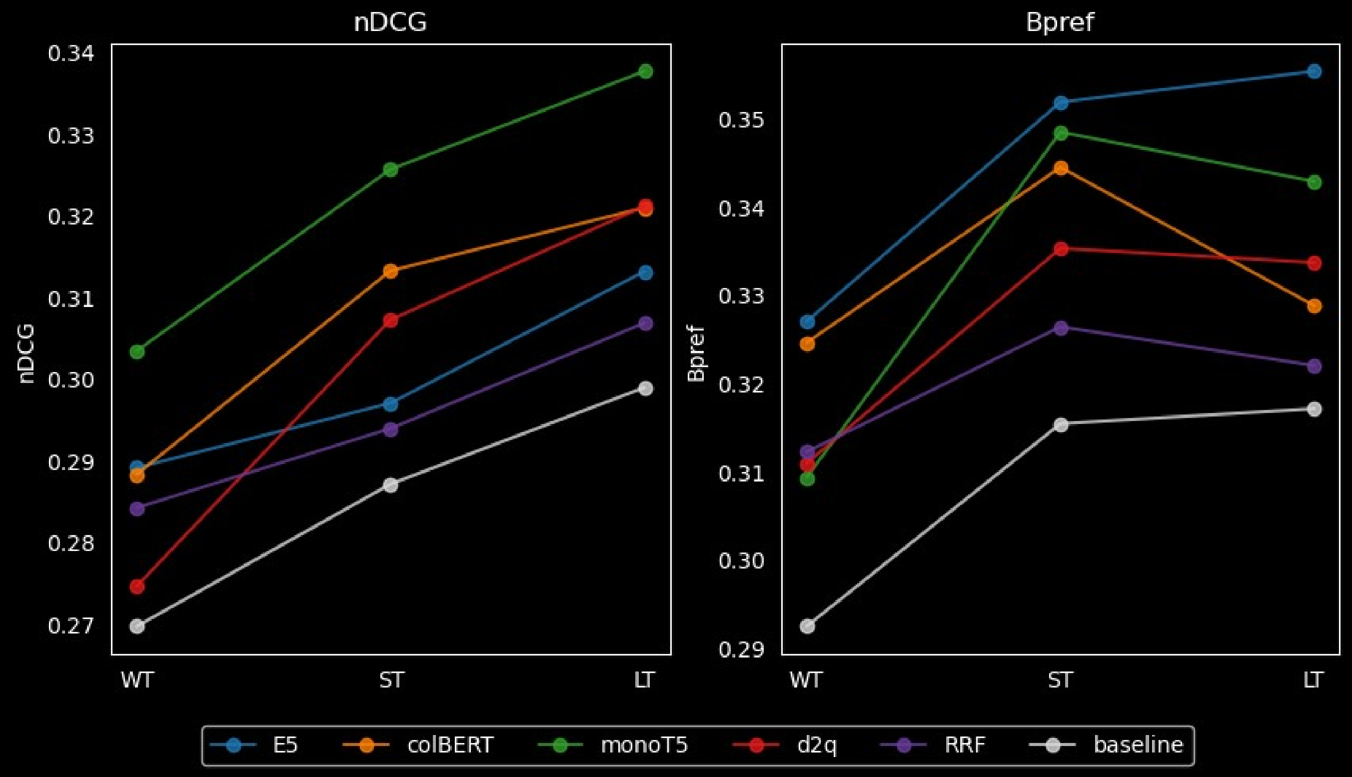
Slightly underpowered results

		ARP			$\mathcal{R}_e\Delta$	
		WT	ST	LT	WT, ST	WT, LT
Bpref	BM25	0.2924	0.3154	0.3171	-0.0230	-0.0247
	RRF	0.3122	0.3264*	0.3220	**-0.0142**	-0.0098
	ColBERT	0.3246	0.3445*	0.3288	-0.0392	-0.0336
	monoT5	0.3093	0.3485*	0.3429*	-0.0244	-0.0228
	d2q	0.3109	0.3353*	0.3337*	-0.0199	**-0.0042**
	E5	**0.3270**	**0.3519***	**0.3554***	-0.0249	-0.0284
P@20	BM25	0.0648	0.0658	0.0722	-0.0010	-0.0074
	RRF	0.0658	0.0657	0.0738	**0.0001**	-0.0080
	ColBERT	0.0704	0.0705*	0.0775*	0.0013	-0.0075
	monoT5	**0.0781***	**0.0768***	**0.0856***	-0.0021	-0.0109
	d2q	0.0684	0.0705*	0.0793*	**-0.0001**	-0.0071
	E5	0.0673	0.0652	0.0726	0.0021	**-0.0053**
nDCG	BM25	0.2697	0.2871	0.2989	-0.0174	-0.0292
	RRF	0.2842*	0.2939*	0.3068*	-0.0097	**-0.0226**
	ColBERT	0.2883	0.3132*	0.3209*	-0.0222	-0.0342
	monoT5	**0.3034**	**0.3256***	**0.3376***	-0.0326	-0.0465
	d2q	0.2746	0.3072*	0.3211*	-0.0249	-0.0326
	E5	0.2891	0.2970	0.3131	**-0.0079**	-0.0240

Effectiveness improves

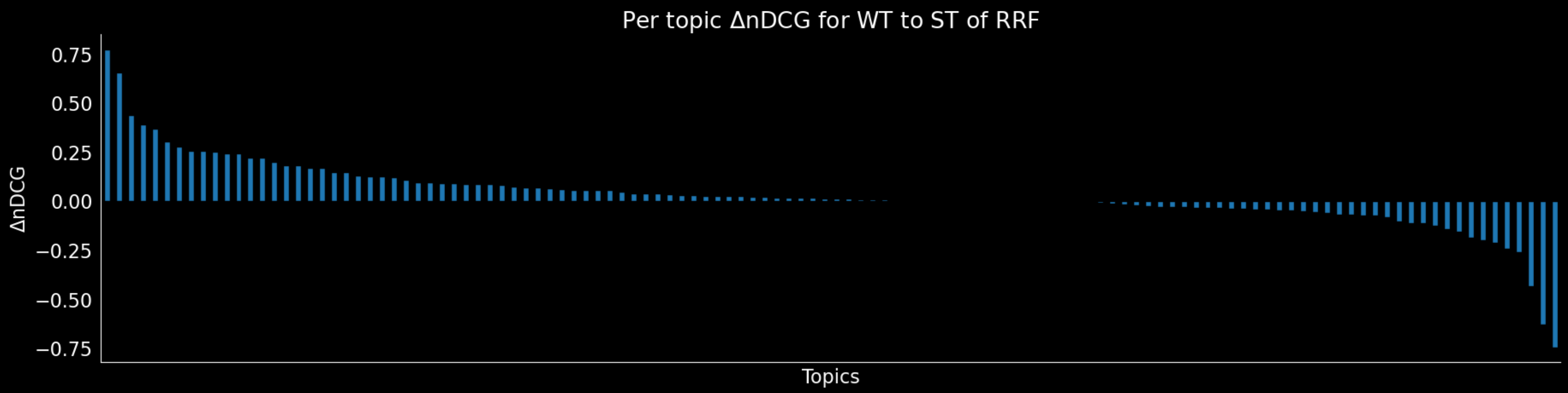
Not a single “best system“

Longitudinal Evaluations



- Effectiveness improved over time
- Consistency deteriorates
- ROS variates across sub-collections and measures

Replicability



- Dynamic Evaluation Environment (EE)
- More detailed evaluations with replicability measures
- Isolate changes and their influence on the effectiveness

Replicability

Measure effects in relation to a Pivot system

$$\mathrm{ER}(\Delta' M^{EE_2}, \Delta M^{EE_1}) = \frac{\overline{\Delta' M^{EE_2}}}{\overline{\Delta M^{EE_1}}} = \frac{\frac{1}{n_{EE_2}} \sum_{j=1}^{n_{EE_2}} \Delta' M_j^{EE_2}}{\frac{1}{n_{EE_1}} \sum_{j=1}^{n_{EE_1}} \Delta M_j^{EE_1}}$$

- **Effect Ratio (ER):** Improvement recovered

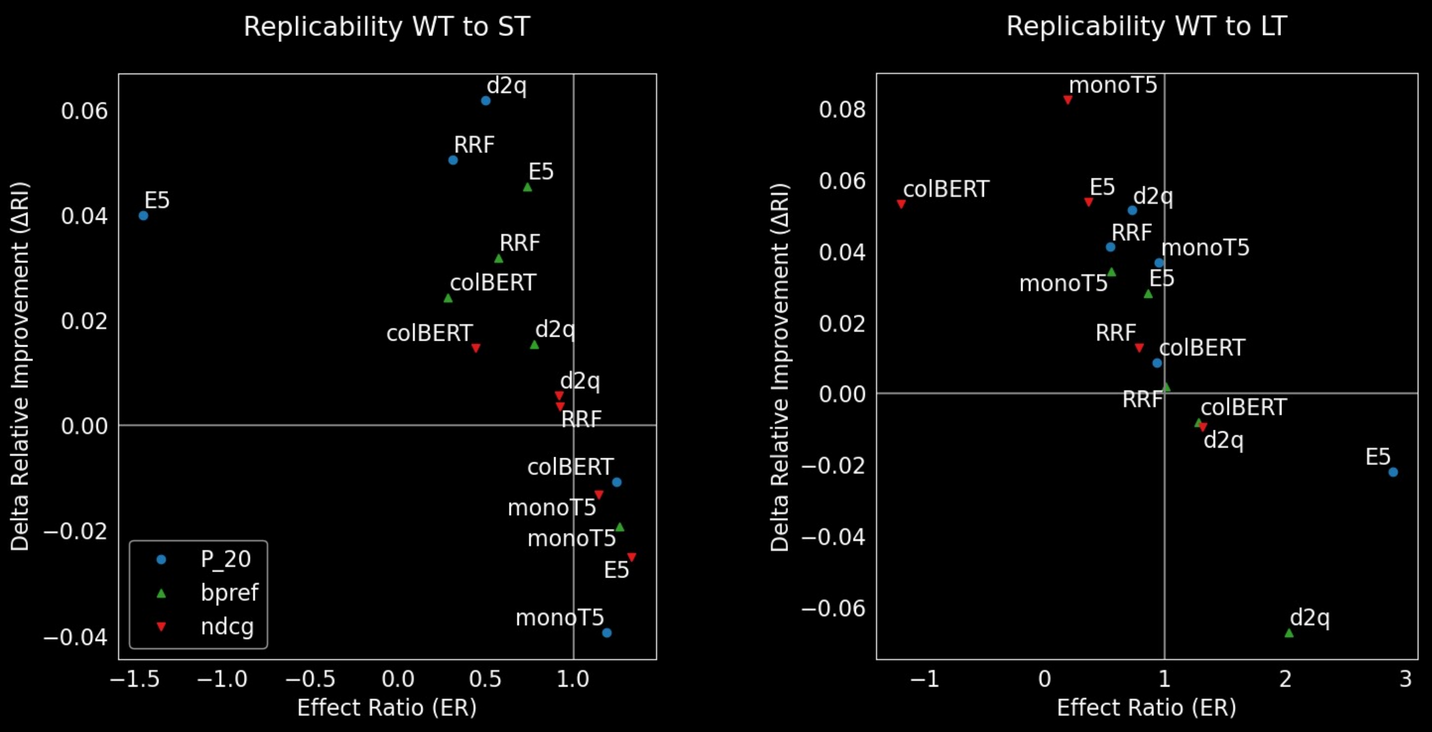
$$\Delta\mathrm{RI} = \mathrm{RI} - \mathrm{RI}'$$

- **Delta Relative Improvement (DeltaRI):** Overall effectiveness recovered

$$\mathrm{RI} = \frac{\overline{M^{EE_1}(S)} - \overline{M^{EE_1}(P)}}{\overline{M^{EE_1}(P)}}$$

$$\mathrm{RI}' = \frac{\overline{M^{EE_2}(S)} - \overline{M^{EE_2}(P)}}{\overline{M^{EE_2}(P)}}$$

Replicability



- Improved absolute scores and replicated relative effect
- Reduced absolute scores and weaker relative effect
- Replication deteriorates over time (shift to right top)

Conclusion

- Effectiveness measured by different measures or for different topics does not necessarily agree with each other.
- Replicability measures seem to be a beneficial addition to gaining further insights.
- Interpretation remains difficult, effects overlay
- Narrowing down the effects by reducing changes in the EE
- Results seem to correlate with ARP and result deltas

Thank You

Questions?

jueri

@juerikeller

jueri.keller@th-koeln.de